

RICE UNIVERSITY

**A Simulation-based Approach to Study Rare Variant  
Associations Across the Disease Spectrum**

by

**Rosa C. Banuelos**

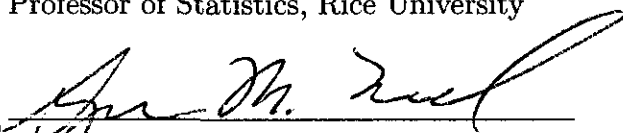
A THESIS SUBMITTED  
IN PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE

**Doctor of Philosophy**

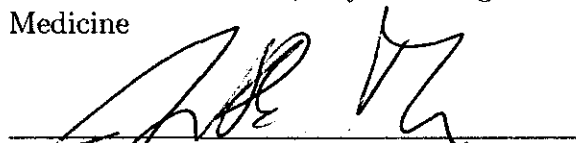
APPROVED, THESIS COMMITTEE:



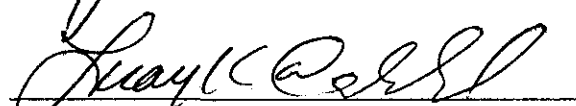
Marek Kimmel, Committee Chairman,  
Professor of Statistics, Rice University



Suzanne M. Leal, Professor of Molecular  
and Human Genetics, Baylor College of  
Medicine



James R. Thompson, Noah Harding  
Professor of Statistics, Rice University



Luay K. Nakhleh, Associate Professor of  
Computer Science, Rice University

HOUSTON, TEXAS

MAY, 2013

## Abstract

# A Simulation-based Approach to Study Rare Variant Associations Across the Disease Spectrum

by

Rosa C. Banuelos

Although complete understanding of the mechanisms of rare genetic variants in disease continues to elude us, Next Generation Sequencing (NGS) has facilitated significant gene discoveries across the disease spectrum. However, the cost of NGS hinders its use for identifying rare variants in common diseases that require large samples. To circumvent the need for larger samples, designing efficient sampling studies is crucial in order to be able to detect potential associations. This research therefore evaluates sampling designs for rare variant-quantitative trait association studies. More specifically, a statistical framework is presented such that designs based on selective sampling are properly accounted for. This work also assesses the effects on power that sampling individuals from existing public cohorts can have on the design. Performing simulations and evaluating common and unconventional sampling schemes results in several noteworthy findings. Specifically, the extreme-trait design is the most powerful design for analyzing quantitative traits. This research shows that sampling more individuals from the extreme of clinical interest does not increase power.

Sampling design can have a greater role in gene discovery for monogenic diseases as the focus moves away from family data to population-based data, in part due to the

advancements in NGS. These advances have facilitated approaches based on variant filtering, which have served as a “proof-of-concept” approach for the discovery of disease-causing genes in Mendelian traits. Still formal statistical methods have been lacking in this area. However, combining variant filtering schemes with existing rare variant association tests is a practical alternative. In this work, a variant filtering step is implemented prior to performing rare variant association testing in Mendelian traits. Specifically, six burden-based rare variant tests are evaluated in the presence of genetic heterogeneity and genotyping errors. This research shows that with low locus heterogeneity, these tests are powerful for testing association. With the exception of the weighted sum statistic (WSS), the remaining tests were very conservative in preserving the type I error when the number of affected and unaffected individuals was unequal. The WSS, on the other hand, had inflated type I error as the number of unaffected individuals increased.

The framework presented can serve as a catalyst to improve sampling design and to develop robust statistical methods for association testing.

## Acknowledgements

First, I thank my family for supporting me in this endeavor, in particular my mom and my brother Felix for their constant encouragement and love from the very beginning. I am grateful to my husband for going through this process with me and my son for coming along at the right time to become my motivation to continue.

I would like to thank Dr. Marek Kimmel, my committee chair, for his mentoring during critical periods in my training and for serving as a liaison when needed. I am grateful to my committee members Drs. Jim Thompson and Luay Nakhleh for their patience, support, and compassion. My sincere thanks to Dr. Suzanne Leal for providing me with the opportunity to work on several projects in her lab and for the important lessons on maintaining professionalism, cooperation and transparency. In addition, I thank Dajiang and Gao for all of their support during my time in the Leal lab.

I am indebted to Dr. Richard Tapia and Theresa Chatman for their mentoring and support throughout my time at Rice. To all my friends who would listen to all the good and bad, who reviewed my work, or simply joined me for a coffee break: Andria, Beibei, Debbie, Erin, Kalatu, Paula, Regie, . . . , and many more, Thank You!

Finally, my dissertation work was funded under a training fellowship from the Keck Center of the Gulf Coast Consortia, on the Training Program in Biomedical Informatics, National Library of Medicine (NLM) T15LM007093 and the Alliances for Graduate Education and the Professoriate (AGEP) Program by the National Science Foundation Cooperative Agreement HRD-0450363.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>Abbreviations</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Background and Literature Review</b>	<b>4</b>
2.1 Competing Hypothesis on Common Disease Etiology . . . . .	6
2.2 Next Generation Sequencing (NGS) . . . . .	8
2.2.1 Current Sequencing Technologies . . . . .	8
2.2.2 Exome Sequencing . . . . .	9
2.2.3 Limitations of NGS and Exome Sequencing . . . . .	11
2.3 Rare Variant Association Methods . . . . .	12
2.3.1 Burden Tests . . . . .	13
2.3.2 Directional Tests . . . . .	21
2.4 Methods for Selective Sampling . . . . .	25
2.5 Gene Discovery Approaches for Mendelian Traits . . . . .	31

2.6	Difficulties in the Quest for Rare Variants . . . . .	36
<b>3</b>	<b>Selective Sampling in Rare Variant Association Studies</b>	<b>39</b>
3.1	Introduction . . . . .	39
3.2	Methods . . . . .	42
3.2.1	Sampling Framework . . . . .	43
3.2.2	Description of Study Designs and Sample Selection . . . . .	44
3.2.3	Approach for Analyzing Dichotomized Phenotypes . . . . .	48
3.3	Simulation Set-Up . . . . .	48
3.4	Results . . . . .	50
3.4.1	Evaluation of Type I Error . . . . .	50
3.4.2	Effect of $\beta$ and Percent Causal Variants on Power . . . . .	51
3.4.3	Effect of Sample Size and Threshold on Power . . . . .	54
3.4.4	Effect of Asymmetric Thresholds and Unbalanced Sampling on Power . . . . .	56
3.5	Discussion . . . . .	61
<b>4</b>	<b>On the Application of Rare Variant Association Tests to Mendelian Traits</b>	<b>64</b>
4.1	Introduction . . . . .	65
4.1.1	On Traditional Linkage Analysis . . . . .	66
4.1.2	On the use of NGS . . . . .	66
4.1.3	Exome Sequencing Coupled with Filtering Schemes . . . . .	67
4.2	Methods . . . . .	68
4.2.1	Filtering Implementation . . . . .	68
4.2.2	Burden Tests . . . . .	69
4.3	Simulation Set-up . . . . .	70
4.3.1	Genetic Data . . . . .	70
4.3.2	Scenarios Considered . . . . .	71

4.4	Results . . . . .	72
4.4.1	Evaluation of Type I Error . . . . .	72
4.4.2	Power Analysis . . . . .	75
4.5	Discussion . . . . .	81
<b>5</b>	<b>Conclusion and Future Work</b>	<b>84</b>
	<b>Bibliography</b>	<b>87</b>

# List of Figures

2.1	Disease Spectrum of Hypertriglyceridemia . . . . .	5
2.2	Illustration of Exons in a Gene . . . . .	11
2.3	Illustration of the Extreme-Trait Design . . . . .	27
2.4	Illustration of the One-Tail Design . . . . .	27
2.5	The Almost-Extreme Sampling Design . . . . .	31
2.6	Filtering Pipeline for Variant Prioritization . . . . .	34
3.1	Illustration of Three Sampling Designs . . . . .	46
3.2	Power versus Percent Causal Variants, Balanced Extreme Trait Design	52
3.3	Power versus Percent Causal Variants, Balanced One-tail Design . . .	53
3.4	Power versus Percent Causal Variants, Balanced One-tail Design with a Random Sample . . . . .	53
3.5	Power Comparison for Thresholds $\Phi^{-1}(0.95)$ and $\Phi^{-1}(0.90)$ for an Ex- isting Cohort . . . . .	54
3.6	Power Comparison of $\beta = 0.5$ and $\beta = 1.0$ for an Existing Cohort . . .	55
3.7	Power Comparison of Three Designs for Varying Sampling Ratios . .	56
3.8	Power under the Effects of Asymmetric Thresholds and Unbalanced Sampling for an Existing Cohort . . . . .	58
4.1	Filtering Scheme of Variants for Mendelian Traits . . . . .	69
4.2	Power versus Number of Cases . . . . .	78
4.3	Power versus Proportion of Locus Heterogeneity, Unbalanced Sampling	80



# List of Tables

2.1	Comparison of Sanger and NGS . . . . .	10
2.2	$2 \times 2$ Contingency Table . . . . .	14
3.1	Description of Sampling Designs Implemented . . . . .	45
3.2	Simulation Information for Africans . . . . .	48
3.3	Type I Error Comparison for Balanced Designs . . . . .	50
3.4	Type I Error under Unbalanced Sampling for an Existing Cohort . . .	50
3.5	Type I Error Comparison for Varying Thresholds and Unbalanced Sam- pling, $\alpha = \{0.05, 0.001\}$ . . . . .	51
3.6	Power under Varied Sampling Ratios, $\beta$ 's, and Percent Causal Variants	57
3.7	Power under Balanced Sampling and $q_U = 0.90$ , Varied $q_L$ and $\beta$ . . .	60
4.1	List of Six Burden-Based Association Tests Implemented . . . . .	70
4.2	Simulation Information for Europeans . . . . .	70
4.3	Noise Introduced in Simulations . . . . .	71
4.4	Information on Sample Sizes and Scenarios . . . . .	72
4.5	Type I Error Comparison for a Dominant Trait under Balanced Sampling	73
4.6	Type I Error Comparison for a Dominant Trait under Unbalanced Sam- pling, $\alpha = 0.05$ . . . . .	74
4.7	Type I Error Comparison for a Dominant Trait under Unbalanced Sam- pling, $\alpha = 0.01$ . . . . .	75
4.8	Power of the CMC under Varied Levels of Genetic Heterogeneity . . .	76

4.9	Power Comparison of Six Burden Tests under Varied Locus Heterogeneity	79
4.10	Power Comparison of Six Burden Tests under Varied Proportion of Causal Variants . . . . .	79
4.11	Power Comparison of Six Burden Tests under Unbalanced Sampling and Varied Proportion of Causal Variants . . . . .	81

## Abbreviations

<b>CD/CV</b>	Common Disease-Common Variant Hypothesis
<b>CD/RV</b>	Common Disease-Rare Variant Hypothesis
<b>CNV</b>	Copy Number Variation
<b>dbGaP</b>	Database of Genotypes and Phenotypes
<b>dbSNP</b>	Single Nucleotide Polymorphism Database
<b>DT</b>	Dichotomized Trait
<b>GWAS</b>	Genome-wide Association Study
<b>HWE</b>	Hardy-Weinberg Equilibrium
<b>IBD</b>	Identical-By-Descent
<b>LD</b>	Linkage Disequilibrium
<b>LOD</b>	Log Odds Score
<b>LOF</b>	Loss-of-Function
<b>MAF</b>	Minor Allele Frequency
<b>NGS</b>	Next-Generation Sequencing
<b>NHGRI</b>	National Human Genome Research Institute
<b>NS</b>	Nonsynonymous
<b>OMIM</b>	Online Mendelian Inheritance in Man
<b>QT</b>	Quantitative Trait
<b>QTL</b>	Quantitative Trait Locus
<b>SNP</b>	Single Nucleotide Polymorphism
<b>UDP</b>	Undiagnosed Disease Program from NIH

# Chapter 1

## Introduction

Advances in Next generation sequencing (NGS) technologies have advanced our knowledge of many Mendelian and common diseases, in particular understanding the role of rare genetic variants in the etiology of disease. However, it is still relatively expensive to generate sequence data at the large scale required for many statistical applications. To circumvent the need of larger samples for common diseases with an underlying quantitative trait (QT), two strategies can be considered in the study design: sampling individuals with extreme trait values and utilizing publically available phenotyped cohorts. Analyzing the extremes of QT distributions was described earlier by Lander and Botstein (1989) for animal breeding studies. Lander and Botstein (1989) reported that selecting the progeny with extreme QTs provided most of the information for linkage [50]. More recently, Cohen et al. (2004) resequenced genes in individuals in only the upper and lower tails of the LDL cholesterol distribution and showed that individuals in each of the tails harbored different rare genetic variants than those in the opposite tail [17]. Freely available public cohort databases that contain thorough phenotypic and genotype information on thousands of individuals is also a viable approach to acquire large samples as these individuals could be part of the comparison group or serve as controls.

New rare variant association methods and spin-offs of existing ones continue to emerge at a fast pace for analyzing complex traits. Yet a very limited number of methods beyond traditional linkage analysis for Mendelian traits exists, even though more than half of the culprit genes for such traits remain unknown. The difficulties of gene mapping in Mendelian traits result because such traits are very rare and there is a need to obtain sufficient families to observe cosegregation. A successful approach to finding disease-causing genes has been through resequencing or targeted resequencing of genetic regions such as the exome in a single affected individual or even a handful of individuals. Exome sequencing coupled with variant filtering approaches that reduce the list of probable disease-causing genes has produced significant genetic discoveries, as is the case for the *MLL2* gene in Kabuki syndrome [73] and the *DHODH* gene in Miller syndrome [74]. However, such variant filtering approaches are regarded as “proof-of-principle” without formal statistical methods to back them up.

## Thesis Outline

The remainder of this thesis is divided into four chapters. Chapter 2 provides a background and literature review of the major advancements in NGS and statistical methods developed for rare variant association testing in both common and Mendelian diseases. Specifically, I discuss two predominant classes of association tests for complex traits: burden-based and directional tests. I also discuss selective sampling for quantitative traits. I close chapter 2 by discussing traditional and recent gene mapping approaches for Mendelian traits and overall difficulties in finding rare genetic variants.

In chapter 3, a statistical framework is presented for selective sampling designs for rare variant-complex QT associations. By performing thorough simulations, type I error and power are evaluated for different sampling schemes from the general population and fixed-size cohorts. The primary finding is that the extreme-trait design

is the most powerful design, with only a mild type I error inflation, within an  $\varepsilon$  of  $\alpha$ . In addition, when carrying out selective sampling from an existing cohort, selecting more individuals from the tail of clinical interest does not increase power.

Chapter 4 evaluates the performance of six burden-based rare variant association tests for rare-variant Mendelian-trait association after a variant filtering step. Specifically, the type I error and power of the test are formally assessed via a simulation study of both dominant and recessive traits with different levels of genetic heterogeneity. To allow for noise from the sequencing and genotyping process, the proportions of causal and neutral variants that are missing are also varied. When including an equal number of affected and unaffected individuals, the six tests conservatively preserve the type I error. Finally, the tests were sufficiently powerful when a small fraction of the causal variants are missing and there is low locus heterogeneity.

In chapter 5, I provide a summary of the major contributions of this thesis and discuss future research that can stem from my dissertation work. I also discuss some of the difficulties in association testing for both Mendelian and complex traits.

## Chapter 2

# Background and Literature Review

The spectrum of human traits ranges from Mendelian to multifactorial traits. Mendelian traits usually result from a mutation in a single gene and are often qualitative with a clear cut outcome (present/not present). Multifactorial traits take on qualitative or quantitative values and are influenced by a mix of genetic and environmental factors and possibly interactions among them. There is no clear boundary for separating single-gene disorders from multifactorial traits [57] and there are multifactorial traits with monogenic forms. Figure 2.1 illustrates this point for hypertriglyceridemia (HTG), a metabolic lipid disorder, that can have monogenic or polygenic forms based on the levels of plasma triglycerides (TG) [42].

Because the makeup of multifactorial traits is more involved, these type of traits are better known as *complex traits*. Examples of such traits include height, body mass index (BMI), and cancer. Complex traits that are of primary public health interest are common diseases in the general population like cardiovascular disease (CVD) which has high mortality and elevated health care cost in the U.S. [68]. Unraveling the makeup of common diseases, in particular their genetic factors, continues to be a high priority to decrease the public health burden.

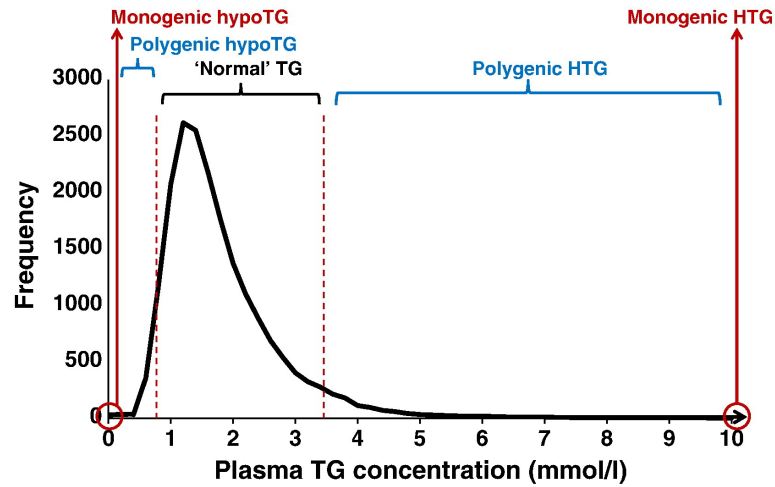


Figure 2.1: Image of the disease spectrum of HTG. This image illustrates that the most extreme levels of triglycerides can lead to Mendelian-forms of HTG. Reprinted with permission from Elsevier: BBA-Molecular & Cell Biology of Lipids, 1821(15), 833-42, copyright (2012) [42].

In the past decade, genome-wide association studies (GWAS) have been the leading approach for searching for culprit genes in common diseases. GWAS are often metaphorically referred to as “fishing expeditions” [44] because the entire genome is scanned for genetic variants without *a priori* knowledge of a candidate gene. Recent GWAS are able to interrogate millions of Single Nucleotide Polymorphisms (SNPs) in thousands of individuals. One of the most notable successes of GWAS came in 2005 with the reported association between complement factor H (*CFH*) gene on chromosome 1 and age-related macular degeneration (MD) [45]. The association between *CHF* and MD is of particular interest because it required only 96 individuals and this association has been replicated, which are both unusual for GWAS. As of July 16, 2012 the National Human Genome Research Institute (NHGRI) has reported more than 1,300 papers on associations between 6,596 SNPs and more than eighty traits, many of which are replication studies [35].



## 2.1 Competing Hypothesis on Common Disease Etiology

GWAS are driven primarily by the Common Disease/Common Variant (CD/CV) hypothesis, which states that the variability in complex traits result from variants with small to moderate effect size that are common in the population [49]. These common variants have a minor allele frequency (MAF) of 1% [12] or more in the general population. The classical example used to support CD/CV is the association of Alzheimer’s disease and the Apolipoprotein E (*APOE*) gene. The frequency of the *APOE*- $\epsilon$ 4 allele in the U.S. ranges between 15-30% and is associated with 50% of Alzheimer cases [93].

Based on twin studies, many common diseases have a heritability between 40% and 90% [58] yet the majority of common variants identified by GWAS have not been able to explain more than 20% of the genetic variability in many complex diseases [62]. Many of these common genetic variants have been associated with the trait under study, yet they often are not the true causal variant, i.e., the variant that alters the function of a gene and in turn alters the phenotype. If the discovered variant is causal, it may be necessary but alone is not sufficient to alter the trait outcome. The reason is that common variants typically have small effect sizes with odds ratios (ORs) no greater than 1.5 [61].

This has lead to a search for more contributing factors to the remaining proportion of genetic variability or “missing heritability” of complex traits [60]. One option has been to focus on uncommon variants, in particular to rare variants. The increased interest in the role of rare variants results from an alternative hypothesis to the CD/CV, the Common Disease/Rare Variant (CD/RV) hypothesis. CD/RV states that many variants with MAF less than 1% underlie common disease [87], and thus explain a

greater proportion of the genetic variability. These variants alone may be rare but as a group they can be common in the population. Since rare variants are more likely to be nonsynonymous [12] and functional [31], it is believed that they will have a greater effect on common diseases.

While the study of rare variants is not new, the rekindled interest in them has amassed several findings. The association between high density lipoprotein cholesterol (HDL-C) and multiple rare variants in the *NPC1L1* gene [17] is one of the first to be referred to. Another successful association includes one reported between Type 1 Diabetes (T1D) and rare variants in the *IFIF1* gene [70].

A recent conjecture suggests that the signal that has been picked up by GWAS at loci with common variants likely results from a rare variant nearby creating what is called a *synthetic association* [20]. The synthetic association hypothesis is plausible because rare and common variants can reside on the same loci. However, this hypothesis is still being debated. In addition, GWAS have produced notable gene discoveries as discussed earlier.

Rare variants are not amenable to detection by the technology used for common variants. For common variants, there exists comprehensive databases for selecting tagSNPs based on the results from projects such as the HapMap, which were conducted to catalogue human variation with a MAF of at least 1%. Such catalogues were key to developing reliable SNP chips to detect common variants. While projects are underway to create similar databases for rare variants, alternative detection technology will be necessary.

## 2.2 Next Generation Sequencing (NGS)

Tagging markers based on linkage disequilibrium (LD), i.e. allelic association, as done for GWAS is not suitable for rare variants because it is unlikely that more than one rare variant resides on the same haplotype. Rare variants must be searched for directly in an unbiased manner using sequencing technology. Sequencing refers to determining the exact DNA base (A, T, G, or C) ordering in a genetic region. In 1977, Frederick Sanger developed an efficient sequencing approach based on a dideoxy terminator technique [92]. The ABI capillary sequencing machines based on the Sanger method have been widely used. Sanger sequencing had been the primary sequencing approach until the mid-2000's. Next Generation Sequencing (NGS) technology emerged in 2005 allowing for massive parallel sequencing [30] in less time and a lower cost than Sanger sequencing. The first human genome sequenced using NGS took  $\sim 4$  months and cost about \$1 million in comparison to the four years and \$100 million required to complete the first human genome using the Sanger method [89].

### 2.2.1 Current Sequencing Technologies

NGS or what is now called *second-generation sequencing* technology has enabled high-throughput sequencing of whole human genomes. Sequencing a human genome, which consists of  $\sim 3.2 \times 10^9$  bp [29], using NGS takes an average of one week [96]. The leading companies for NGS include Roche, Illumina and Applied Biosystems (ABI). Roche's and Illumina's sequencing technologies rely on synthesis-based sequencing. The SOLiD<sup>a</sup> by ABI is based on a ligation method. More recent sequencing technology, or *third-generation sequencing* methods are based on single-molecule sequencing techniques [13] and are currently developed by Helicos BioSciences, Pacific Biosciences, Oxford Nanopore and Ion Torrent. Single molecule sequencing was originally introduced by Helicos BioSciences [34]. While Oxford Nanopore also uses single-molecule sequencing, its goal is to "develop the first label-free, single molecule

DNA sequencing technology” [80]. Table 2.1 provides a list of common sequencing technologies and key features of the employed sequencing instruments. In summary, 454 gives greater read length and is the most expensive. While Illumina and AB are most cost-effective, they are more error prone due to their shorter reads [13].

As of October 2012, NHGRI reports that sequencing the entire human genome is estimated to cost \$6,618<sup>1</sup> [101]. While the cost has dramatically decreased in comparison to the first genome sequenced using NGS, large scale studies are infeasible because of the large sample sizes required to detect low frequency variants. Alternatives to whole-genome sequencing include targeted sequencing or re-sequencing of selected genes and other regions of the genome.

## 2.2.2 Exome Sequencing

A gene is made up of a 5'-untranslated region (5'-UTR), introns, exons, and 3'-UTR as illustrated in figure 2.2. There are ~20,000-25,000 genes in the human genome. Of particular importance are the protein-coding regions of genes, or exons, because these regions are more likely to harbor functional rare variants and more than 85% of disease-causing variants are located in this region [16]. All exons comprise about 1 to 2% of the human genome and collectively all exons are dubbed the *exome*. The exome contains ~180,000-200,000 exons. In any one individual there can be between 15,000 to 20,000 variants per exome [75], with estimates increasing to 24,000 in African Americans [6].

Exome sequencing, or sequencing only the exons and flanking introns, is carried out by using reagent kits like Illumina's TruSeq, Nimblegen SeqCap EZ exome or Agilent's SureSelect Target Enrichment kits on the platforms discussed in the previous section. Exome sequencing offers a feasible alternative to whole-genome sequencing because

---

<sup>1</sup>This estimate reflects the “production costs” as described at [www.genome.gov/sequencingcosts](http://www.genome.gov/sequencingcosts)

Generation	Year	Method	Company/Instrument	Read Length (bp)	Throughput/Time
1st	1977	Dideoxy/Capillary	Sanger on ABI 3730XL	500-800	$1.2 \times 10^3$ bp/day
2nd	2005	Pyrosequencing	Roche 454 GS-FLX	250-1000	400 Mb/day
	2006	Synthesis	Illumina HiSeq 2000 Series	200-250	$\sim 1500$ Mb/day
	2008	Ligation	SOLiD Systems <sup>a</sup>	25-35	7 to $> 20$ Gb/day <sup>b</sup>
3rd	2008	Single Molecule	Helicos Biosciences HeliScope™ [34]	25-55	$\sim 1$ Gb/day
	2010	Single Molecule	Pacific Biosciences PacBio RS [81]	$< 2,200$	100 Mb/2h run
	2010	Nanopore	Oxford Nanopore GridION [80]	$< 10,000$	10 Gb/day
	2010	Single Molecule	Ion Torrent PGM/Proton [38]	$< 200$	100 Mb to $> 1$ Gb/day

Table 2.1: Comparison of Sequencing Methods and Technology [13, 30, 96].

<sup>a</sup>SOLiD stands for “Sequencing by Oglio Ligation and Detection”

<sup>b</sup>Throughput by Model: 7-9 Gb/Day for 5500 and 10-15 Gb/day or  $> 20$ /day for 5500xl, see [4].

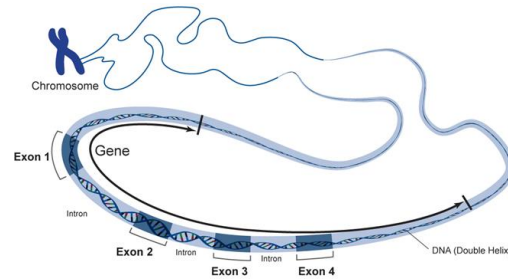


Figure 2.2: Image of exons and introns in a gene. National Institutes of Health. National Human Genome Research Institute. “Talking Glossary of Genetic Terms.” Retrieved July 24, 2010, from <http://www.genome.gov/glossary> [76].

its lower sequencing cost translate to an increase in the sampling potential by up to ten-fold. Exome sequencing in as little as one to ten individuals has yielded novel gene discoveries. By sequencing the exomes of four individuals, Ng et al. (2009) reported the discovery of causal variants in the *DHODH* gene responsible for Miller syndrome [74] and shortly after this group also reported the identification of the *MLL2* gene causing Kabuki syndrome [73].

### 2.2.3 Limitations of NGS and Exome Sequencing

Sequencing technology would ideally cover all variant types at high coverage and almost error free [14]. Unlike Sanger sequencing, NGS has a greater error rate and there is still no consensus on targets or genotype calling [6, 98]. While most sequencing systems have an accuracy rate above 99%, newer technology such as Oxford Nanopore’s have an accuracy of 96% [80]. Many discoveries from NGS are validated using Sanger sequencing. To do away with such a validation step would require NGS to reach the high fidelity obtained by microarray chips. The high genotyping call rates of genotyping chips makes them vital in gene discovery and may often be preferred over exome or whole genome sequencing [28]. Recently, NIH’s Undiagnosed Disease Pro-

gram (UDP) showed that genotyping chips were pivotal in diagnosing rare conditions in which a homozygous state or recessive mode of inheritance was suspected [27]. For the purpose of detecting rare variant associations with complex traits, genotyping chips will not be an option until rare variation is properly catalogued. NGS poses a large financial constraint that is expected to be alleviated with the coming of the \$1,000 genome. Until then, exome sequencing will remain a viable alternative. While exome sequencing reduces costs, it is primarily limited because it only covers  $\sim 1-2\%$  of the entire genome and up to 10% of the exome can be missed or not properly covered [6]. Non-coding regions can also harbor important functional variants. For example, variants in the introns of gene *FGFR2* are involved in gene regulation and are associated with increased risk for breast cancer [22]. In addition to the limitations posed by technology, there is still a dire need for statistical methodology and sampling designs to link actual genetic discoveries to phenotypes.

## 2.3 Rare Variant Association Methods

Rare variants by definition have a low allele frequency which makes population-based methods difficult to implement. In the past five years many different flavors of rare-variant association methods have been introduced, ranging from contingency table-based tests to more complicated methods such as Bayesian Hierarchical models. Here, I present the tests that serve as the foundation to many of newer tests or tests that are commonly used. The primary tests used for rare variant association can be broadly classified as Burden (i.e., collapsing) or Directional tests. Burden tests assess the number of rare variants carried in affected individuals relative to the number in controls. Directional tests are those that account for whether a variant has a risk or a protective effect. For both type of tests, there can be a weighting mechanism that accounts for variant enrichment in cases or that weighs based on functional information about the variant. Below, I introduce notation that will be consistent for the methods

described in the following sections. I assign the subscript  $i$  for individuals  $\{1, \dots, n\}$ ,  $j$  for variants  $\{1, \dots, m\}$ , and  $k$  for groups  $\{1, \dots, l\}$ . The number of affected/case individuals (1) is denoted by  $n_1$  and similarly  $n_0$  for unaffected/control individuals (0).

The quantitative trait is denoted by  $Y$  while binary traits such as affection or case-control status are denoted by  $D$ .  $X$  represents the non-genetic covariates with coefficients  $\beta$ . Genotype information is denoted by  $g_{ij}$  and can take values  $\{0, 1, 2\}$  for the number of alleles individual  $i$  carries at site  $j$ .  $\vec{G}$  is the vector of genotypes or multi-site genotypes, e.g., the vector for individual  $i$  is  $\vec{G}_i = (g_{1i}, g_{2i}, \dots, g_{mi})$  for  $m$  variant sites.  $I_j(i)$  is the indicator function for presence of at least one minor allele in individual  $i$  at site  $j$ . Finally, weights are designated by  $w$ .

### 2.3.1 Burden Tests

Collapsing tests are based on creating groups to analyze rare variants and create “super-variants” [8]. For these types of tests, it is important to determine the best way to collapse variants. Usually this is by gene or MAF.

#### The Cohort Allelic Sums Test (CAST)

The CAST by Morgenthaler and Thilly (2007) was the first method to suggest that mutations be collapsed into groups. CAST is based on comparing the carrier frequencies between affected and unaffected individuals. For a pre-specified grouping criteria, such as gene, the collapsing takes place where an individual carries at least one mutation at a gene. The carrier frequencies between affected and unaffected individuals can be compared using Fisher’s exact test or a Chi-square test [66].

Fisher’s exact test assumes that the marginals of a contingency table are fixed. If we let  $A$  denote the number of affected individuals with one or more rare variants, then for



	0	$\geq 1$	
Affected (1)	$a$	$b$	$n_1$
Unaffected (0)	$c$	$d$	$n_0$
	$a + c$	$b + d$	$n$

Table 2.2:  $2 \times 2$  Contingency Table

a  $2 \times 2$  contingency table (as in table 2.2), the probability of observing  $O = b$  affected individuals with one or more variants can be computed using the hypergeometric distribution:

$$Prob(A = b) = \frac{\binom{n_1}{b} \binom{n_0}{d}}{\binom{n}{b+d}}.$$

The p-value is then calculated as the sum of probabilities of tables where one would observe  $b$  or more affected individuals with one or more variants,  $\sum_{A \geq b} P(A|(b+d))$ .

CAST is straightforward to implement. A limitation of CAST is that the causal variant would not be easily discerned due to the collapsing. In addition, Fisher's exact test is conservative.

### Combined Multivariate and Collapsing (CMC)

Li and Leal (2008) extended CAST to a multivariate setting in the CMC method for detecting rare variants. This method also collapses the genotypes for affected and unaffected individuals as CAST. While typically only rare variants with minor allele frequency (MAF)  $\leq 0.01$  are collapsed, CMC allows for common variants to be in

their own group. Hypothesis testing can be carried out using a multivariate test such as Hotelling's  $T^2$  [53].

The CMC is implemented as follows:

First,  $m$  markers are grouped into  $l$  groups ( $l < m$ ). Let  $\vec{G}_j^1 = (G_{j1}^1, \dots, G_{jn_1}^1)^T$  be the vector of genotypes at site  $j$  for affected individuals. Similarly  $\vec{G}_j^0 = (G_{j1}^0, \dots, G_{jn_0}^0)^T$  is defined for unaffected individuals. The proportion of individuals for each site  $j$  is summarized in  $\bar{G}_j^0 = \frac{1}{n_0} \sum_{i=1}^{n_0} G_{ji}^0$  and  $\bar{G}_j^1 = \frac{1}{n_1} \sum_{i=1}^{n_1} G_{ji}^1$ . All sites for affected and unaffected are summarized in  $\bar{\mathbf{G}}^0 = (\bar{G}_1^0, \dots, \bar{G}_m^0)$  and  $\bar{\mathbf{G}}^1 = (\bar{G}_1^1, \dots, \bar{G}_m^1)$ .

To test all variants simultaneously, Hotelling's  $T^2$  (1931) is used to obtain the test statistic:

$$T^2 = \frac{n_0 n_1}{n} (\bar{\mathbf{G}}^0 - \bar{\mathbf{G}}^1)^T \mathbf{S}^{-1} (\bar{\mathbf{G}}^0 - \bar{\mathbf{G}}^1)$$

with pooled covariance matrix ( $\mathbf{S}$ ):

$$\mathbf{S} = \frac{1}{n-2} \sum_{j=1}^{n_0} (\vec{G}_j^0 - \bar{\mathbf{G}}^0)(\vec{G}_j^0 - \bar{\mathbf{G}}^0)^T + \sum_{j=1}^{n_1} (\vec{G}_j^1 - \bar{\mathbf{G}}^1)(\vec{G}_j^1 - \bar{\mathbf{G}}^1)^T.$$

Hotelling's  $T^2$  is asymptotically distributed as an  $F$  distribution,  $\frac{n-m-1}{m(n-2)} T^2 \sim F_{m, n-m-1}$  [53].

The strengths of the CMC is that power can be computed analytically (has a closed form distribution) and common variants can be included in their own group. The drawback is that there can be a loss of power by collapsing variants with protective and risk effects.

### **Weighted Sum Statistic (WSS)**

Madsen and Browning (2009) proposed the WSS, a groupwise test for rare variant association. Variants are grouped according to gene or some other grouping scheme and weights are assigned based on the variant frequency in controls (0). Each individual is assigned a genetic score based on their variant load. The test statistic is based on the ranking of the genetic score for cases [59].

### Implementation steps for WSS:

1. For each variant  $j$  in a group with  $m_j$  variants, a weight is calculated as:

$$\hat{w}_j = \sqrt{n_j \cdot q_j(1 - q_j)} \text{ where } q_j = \frac{m_j^0 + 1}{2n_j^0 + 2}.$$

For SNP  $j$ ,  $m_j^0$  and  $n_j^0$  represent the number of mutant alleles in controls and the number of controls genotyped for SNP  $j$ , respectively.

2. For each individual  $i$ , a genetic score is calculated across all variants sites:

$$\gamma_i = \sum_{j=1}^M \frac{g_{ij}}{\hat{w}_j}.$$

3. Based on the  $\gamma_i$ 's, individuals are ranked and only the ranks of affected ( $A$ ) individuals are summed:

$$x = \sum_{i \in A} \text{rank}(\gamma_i).$$

4. Affection status is permuted  $k$  times repeating the previous three steps.
5. The ranks of the permuted data are averaged ( $\hat{\mu}_{\text{rank}}$ ) and  $N(0, 1)$  standardized:

$$z = \frac{x - \hat{\mu}_{\text{rank}}}{\hat{\sigma}}.$$

A limitation of the WSS is that asymptotic-based methods are not appropriate for rare variants and permutation-based approaches are computationally expensive.

### **Variable Threshold (VT)**

The VT by Price et al. (2010) is a method to group variants using a variable allele-frequency threshold rather than a fixed threshold. The premise of VT is that there exists an unknown threshold that separates functional variants from neutral and/or non-causal variants. The optimal threshold is selected by maximizing the test statistic over the entire threshold range. VT is implemented in a regression framework where

the trait is a function of genotype scores determined by the variable threshold. VT allows for functional information to be incorporated into a weighting scheme similar to the WSS. Significance can be evaluated using permutation or based on a standard normal approximation [86].

#### VT implementation:

The VT score is denoted by:

$$\sum_{j=1}^m \sum_{i=1}^n w_j C_{ij} y_i, \text{ where } C_{ij} \text{ is the number of alleles at SNP } j.$$

For each threshold  $T$  considered, the  $z$ -score is computed and normalized:

$$z(T) = \frac{\sum_{j=1}^m \sum_{i=1}^n w_j^T C_{ij} (y_i - \bar{y})}{\sqrt{\sum_{j=1}^m \sum_{i=1}^n (w_j^T C_{ij})^2}} \propto N(0, 1).$$

The VT test statistic is taken as the maximum  $z(T)$  over all thresholds  $T$  and is denoted by  $z_{max}$ . Significance is assessed by the ratio of the number of  $z_{max}$  values that exceed the test statistic for the unpermuted data out of all permutations performed.

Functional information can be incorporated through the weights by using the probability that an allele with frequency  $p$  is functional ( $\varphi(p)$ ) or by considering a variant's Poly-Phen score ( $S$ ):

$$w_j = \begin{cases} 1 & \text{for nonsense, etc. variants} \\ 0.5 & \text{for common variants} \\ P(S) & \text{for variant with MAF} < 1\% \end{cases}$$

where  $P(S) = P(S \text{ classifies variant as functional})$  [86].

The VT offers several advantages over other methods. It can be used for qualitative or quantitative traits and it allows for other weighting schemes to be used. For example,

if  $w_j = \frac{1}{\sqrt{p_j(1-p_j)}}$ , then the WSS by Madsen and Browning is recovered. In addition, methods based on a fixed threshold can easily be obtained by defining the VT score as follows:

$$\sum_{j=1}^m w_j C_j = \begin{cases} 1 & \text{frequency of SNP } j \leq T \\ 0 & \text{otherwise.} \end{cases}$$

A limitation of VT is the reliance on functional information because methods used to determine the functionality of variants have been shown to have a high error rate.

### RareCover

RareCover by Bhatia et al. (2010) is a model-free method that attempts to select the subset of variants that has the greatest correlation with the phenotype of interest in a fixed window. RareCover utilizes forward variable selection by including only the most discriminating variants into the subset. The discriminating power of the subset is measured by the ability of the selected variants to differentiate between cases and controls. To avoid the computational demand of checking all possible subsets, a greedy algorithm is used to find the most discriminating variant group that maximizes the test statistic over all variants in region [9].

#### RareCover Steps:

Suppose that a window is defined on a subset  $C$  selected from a group of  $S$  variants and a super variant  $A_C$  (called the *union-variant*) is constructed as:

$$A_C = \begin{cases} 1, & \text{if and only if } I_C > 0 \\ 0, & \text{otherwise.} \end{cases}$$

where  $I_C = \sum_{k \in S} G_k$ . For the different variant subsets, the correlation between the disease status and the union-variant  $A_C$  is computed and denoted by  $CORR(A_C, D)$ .

The RareCover test statistic is the maximum correlation between the disease status

and all possible union-variant constructs:

$$CORR(S, D) = \max_{C \in S} CORR(A_C, D).$$

Significance is assessed as the fraction of the permuted samples whose  $CORR$  value matches or exceeds  $t$ . A stopping rule is put in place once two test statistics no longer exceed a pre-specified threshold  $\tau$ . RareCover is computationally expensive. However, the greedy algorithm introduced reduces the required computations from  $\sim n2^{|S|}$  to  $\sim |S|n$  [9].

### Rare Variant Test (RVT)

The RVT by Morris and Zeggini (2010) is an extension of CAST and CMC for quantitative traits (QTs). The QT is modeled under a linear regression framework where rare variants are aggregated by gene and the QT is then regressed with an indicator function or as the proportion of rare variants an individual carries. The QT is assumed to be  $N(\mu, \sigma^2)$  distributed and the test of association is carried out using a Likelihood Ratio Test (LRT) [67].

To carry out RVT:

Let  $n_{gi}$  be the number of rare variants in the  $i$ th individual and  $r_{gi}$  the variants for which at least one copy of the minor allele is carried. When the QT is modeled as a function of the proportion of rare variants carried, then  $E(y_i) = \alpha + \lambda \frac{r_i}{n_i} + \beta \mathbf{x}_i$ . When the phenotype is modeled as a function of the presence of a rare variant, then  $E(y_i) = \alpha + \lambda I_j(i) + \beta \mathbf{x}_i$ . In both models  $\lambda$  can be interpreted as the expected increase for carrying at least one minor alleles in an affected individual relative to an unaffected individual.

One of the strengths of RVT is that it can easily be extended to case-control studies and covariates can be included in the model. However, as pointed out by the authors, to attain power, variants with a MAF between 1% and 5% must be included [67].

### Kernel-based Adaptive Cluster (KBAC)

The KBAC of Liu and Leal (2010) is a method that combines variant classification and association testing under a unified framework. KBAC models an individual's disease risk as a mixture with two components for causal and non-causal variants. Instead of using each variant individually, KBAC uses continuous adaptive weighting of the multi-site genotypes and selects the weights based on the kernel of the known non-causal variant component. KBAC was specifically designed to deal with variant misclassification and allows for interactions [56].

#### KBAC implementation:

For the multi-site genotype  $G_j$ , the risk is denoted by  $R_j = \frac{n_j^1}{n_j}$  where  $n_j^1$  represents the number of affected (1) individuals and  $n_j$  the combined total of affected and unaffected individuals.  $R_j$ , with probability  $\pi_j$ , can be modeled as a mixture distribution with two components based on the kernels  $\kappa_j$ :

$$R_j \stackrel{D}{\sim} \pi_j \overbrace{\kappa_j^0(R_j)}^{\text{non-causal}} + (1 - \pi_j) \overbrace{\kappa_j^A(R_j)}^{\text{causal}}.$$

It is assumed that  $\kappa_j^0(R_j)$  is known when the multi-site genotype  $G_j$  is non-causal and  $\kappa_j^A(R_j)$  is unknown for causal  $G_j$ . The kernel for the known component,  $\kappa_j^0(\bullet)$  is used to adaptively weigh each of the multi-site genotypes.

To determine the individual weight of each variant, the kernel of the known component can be integrated out as:

$$w_j = \int_0^{\hat{R}_j^0} \kappa_j^0 dr = K_j^0(\hat{R}_j).$$

These weights can then be assigned to those individual who have the  $G_j$  genotype.

The one-sided KBAC statistic is given by:

$$\sum_{j=1}^m \left( \frac{n_j^1}{n_1} - \frac{n_j^1}{n_0} \right) K_j^0(\hat{R}_i).$$

This one-sided test can be interpreted as testing for enrichment of causal variants in affected individuals. KBAC can be implemented for two-sided tests when it is of interest to test for frequency differences in affected and unaffected individuals as:

$$\left( \sum_{j=1}^m \left( \frac{n_j^1}{n_1} - \frac{n_j^0}{n_0} \right) K_j^0(\hat{R}_j) \right)^2.$$

The advantages of the KBAC are that it allows for covariates and interactions to be considered and it is robust to the inclusion of non-causal variants [56]. The major disadvantage of the KBAC is that it requires empirical computation of significance.

### 2.3.2 Directional Tests

Tests in this category allow for protective and risk variants to be analyzed jointly. This type of tests are of particular importance because there are instances in which a gene contains variants that can have opposite effects on the trait of interest. An example of a gene with bi-directional effects is gene *PCSK9* and LDL-C, where different variants in this gene can lead to either elevated or decreased LDL-C [19].

#### The adaptive Sum test (aSum)

The aSum test by Han and Pan (2010) extends the Sum test for common variants by Wang and Elston (2007) to include both common and rare variants. aSum was created for case-control designs and is implemented in a marginal regression framework where the variants are coded in a data-adaptive manner. The aSum attempts to reduce the multiple testing burden by combining the effects of several variants into a single common coefficient [33].

#### Implementation Steps of *aSum*

To select SNPs by allele frequency, a pre-selected cut-off,  $\alpha_0$ , is used. Then:



1. Fit the marginal regression model for each SNP  $j$ : Logit  $Pr(D_i = 1) = \beta_0 + \sum_{j=1}^m G_{ij}\beta_j$  obtaining  $\hat{\beta}_{M,j}$  and p-value  $p_{M,j}$ .
2. For SNPs that satisfy  $\hat{\beta}_{M,j} < 0$  and  $p_{M,j} \leq \alpha_0$ , the SNP codings,  $G_{.j}$ , are changed to  $G_{.j}^* = 2 - G_{.j}$ .
3. Fit the common-effect model

$$\text{Logit } Pr(D_i = 1) = \beta_{c0} + \sum_{j=1}^m G_{ij}\beta_c.$$

The score statistic  $U$  and it's covariance  $V$  are calculated as:

$$\vec{U} = \sum_{i=1}^m (D_i - \bar{D})G_j, \quad \mathbf{V} = \bar{D}(1 - \bar{D}) \sum_{i=1}^m (G_j - \bar{G})(G_j - \bar{G})'$$

$$\text{where } \bar{D} = \frac{\sum_{j=1}^m D_j}{m} \text{ and } \bar{G} = \frac{\sum_{j=1}^m G_j}{m}.$$

4. Permute the original data  $B$  times. For each permuted data  $\{(D_i^{(b)}, G_i)\}$ , repeat steps 1-3 to obtain the null score statistic  $U^{(b)}$ .
  5. Calculate the sample mean and sample variance of  $U^{(1)}, \dots, U^{(B)}$  as  $U_0$  and  $V_0$ .
- The aSum test statistic is:

$$aSum = (U - U_0)V_0^{-1}(U - U_0).$$

The p-value of the test can be obtained by matching the moments of the null distribution to those of the empirical distribution . Alternatively, the p-value can be obtained using permutation. The aSum can be constructed in a way that rare and common variants each have their own group and a two-degree of freedom tests is constructed similarly as done in the common-effect model [33].

A drawback of aSum is the that the test is based on the assumption that the SNPs are in in LD, which may not hold since rare variants are often not in LD. aSum is prone

to lose power for large values of  $\alpha_0$ . In addition, the theoretical null distribution is only an approximation, thus permutation may be necessary. While the type I error can be controlled when permutation is used, this can be computationally intensive.

### **C-alpha (C- $\alpha$ )**

The C- $\alpha$  by Neale et al. (2011) is an extension of a test for overdispersion developed by Neyman and Scott (1966). Neale et al. (2011) apply this test to assess the difference in variants observed in cases versus controls. Overdispersion is assessed by the difference in variant copies in cases and controls rather than differences at the individual level. In addition, this new  $C - \alpha$  was specifically developed to allow for the mixture of variants with opposing effects [69].

#### Implementation of $C - \alpha$ :

When a total of  $n_j$  individuals carries variant  $j$ , if we assume that the variant carrier count in cases ( $n_j^0$ ) has a binomial distribution with common probability  $p_0$ , the test statistic  $T$  under the null can be computed as:

$$T = \sum_{j=1}^m \left[ (n_j^0 - n_j p_0)^2 - n_j p_0 (1 - p_0) \right]$$

with variance

$$Var(T) = c = \sum_{n=2}^{\max n} m(n) \sum_{u=0}^n \left[ (u - n p_0)^2 - n p_0 (1 - p_0) \right]^2 f(u|n, p_0).$$

where  $m(n)$  is the number of variants with  $n$  copies and  $f(u|n, p_0)$  represents the probability of observing  $u$  copies at the  $j$ th variant.

Under the null hypothesis,  $T \sim N(0, 1)$ . Since it is possible that there is an excess observation of singletons (i.e., variants seen only once), these are grouped together as a single observation. C- $\alpha$  can also accommodate different sampling designs by incorporating weights  $w_j$  into the test statistic as:

$$T = \sum_{j=1}^m w_j^{-1} \left[ (n_j^0 - n_j p_0)^2 - n_j p_0 (1 - p_0) \right].$$

The C- $\alpha$  test is sensitive to non-causal variants. However, to circumvent a drop in power, the test could be implemented using the weighted version [69].

Wu et al. (2011) introduced the Sequence Kernel Association Test (SKAT) as a generalization of C- $\alpha$  for QTs using weights based on the sample variant frequencies. SKAT can also be implemented as Madsen and Browning's WSS by using variant frequency weights [102]. Recently, Lee et al. (2012) presented a correlation-based weighted version of SKAT and burden tests for dichotomous traits [52].

### **The Replication-based test (RBT)**

The RBT by Ionita-Laza et al. (2011) was created to allow for the inclusion of variants with opposing direction into a weighted sum-like statistic. The RBT proposes a negative log-based weight for the probability that groups have some number of minor allele counts observed in cases and controls. This probability is then calculated based on a Poisson distribution. RBT is implemented under a case-control setting with an equal number of cases and controls [39].

#### Implementation:

Let  $n_k^{k'}$  denote the number of individuals in a group that have cases with  $k'$  copies and controls with  $k$  copies of some variant and assign a weight  $w_k^{k'}$  to such a group. Then the a weighted sum-statistics can be constructed as

$$S = \sum_{k=0}^{N_r} \sum_{k' > k} n_k^{k'} w_k^{k'}$$

where  $N_r$  is the maximum number of times that a variant is observed in controls. When the number of copies in cases exceeds those observed in controls, Ionita-Laza et al. (2011) proposed data-dependent weights:

$$w_k^{k'} = -\log[p(k, k')] \text{ for } k' > k.$$

With the new weights incorporated in to the WSS, the RBT test statistic is denoted as follows:

$$S = \sum_{k=0}^{N_r} \sum_{k' > k} -n_k^{k'} \log[p(k, k')], \text{ } (S_+ \text{ for risk variants}).$$

Assuming that the variant count follows a Poisson distribution, the probability of observing  $k'$  variants in cases and  $k$  in controls, can be calculated using the Poisson CDF (*ppois*) as:

$$p(k, k') = \text{ppois}(k, \hat{f}) \cdot (1 - \text{ppois}(k' - 1, \hat{f})), \text{ where } \hat{f} = \frac{k + k'}{2}.$$

A two-sided test can be implemented by taking the maximum of the statistics  $S_+$  and  $S_-$ ,  $\max(S_+, S_-)$ , where  $S_-$  is the statistic for protective variants. To test for both risk and protective effects, a test statistic can be constructed as the sum of the one-sided test statistics:

$$S_C = S_+ + S_-.$$

Information about the probability that a variant is functional, denoted by  $p(j)$ , can be incorporated into the test statistic as follows:

$$S = \sum_{k=0}^{N_r} \sum_{k' > k} \sum_{j \in (k, k')} -p(j) \log[p(k, k')].$$

When  $p(j) = 1$  for all variants  $j$ , the regular statistic  $S$  is recovered [39]. RBT is a robust test and suffers only a minor power loss for the directionally. Since a case-control design is used, population structure needs to be properly controlled to avoid an increase in false positives.

## 2.4 Methods for Selective Sampling

Because the genetic etiology of complex traits is much more difficult to decipher, creative ways to detect disease-causing genes have been presented. Several studies have shown that for complex quantitative traits, individuals at the extreme of the trait distribution tend to harbor rare variants. As a way to enrich samples to detect rare

variants, individuals with extreme traits can be ascertained in what can be referred to as *selective sampling*. The underlying idea of selective sampling called “selective genotyping” was discussed by Lander and Botstein (1989) in the context of selecting the most informative animal crosses to detect Quantitative Trait Loci (QTLs) using linkage analysis [50]. Lander and Botstein (1989) proposed reducing the number of individuals genotyped by selecting individuals at the tails of the trait distribution under the premise that these individuals would contribute the most linkage information. They presented an approach to determine the contribution of individuals in the extremes to the total linkage information. For example, they reported that  $\sim 80\%$  of linkage information came from individuals beyond one standard deviation from the mean [50].

Selective sampling can take several forms. The most common is the extreme trait (two-tail) design where individuals at both extremes of the distribution are sampled as in figure 2.3. Without loss of generality (WLOG), suppose that individuals with a trait value beyond the upper threshold  $U$  are the affected. Individuals at the other opposite extreme, i.e. individuals whose phenotype falls below the lower threshold  $L$ , can serve as the comparison group. Recently, Edmond et al. (2012) combined exome sequencing and extreme sampling of the age of onset of individuals susceptible to lung infection and discovered missense rare variants in gene *DCTN4* [23].

Alternatively, individuals from only one tail (as shown in figure 2.4) can be sampled and the comparison can be sampled from the general population or an independent cohort. In the following section, I present several methods that have incorporated the idea of selective sampling.

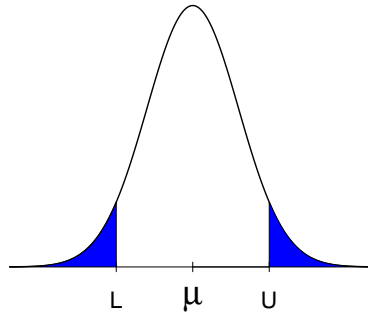


Figure 2.3: Illustration of the Extreme-Trait/Two-Tail Design.

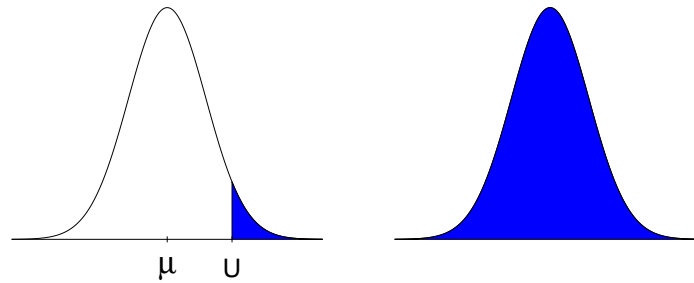


Figure 2.4: Illustration of the One-Tail Design.

### **Rare Variants found Exclusively (RVE)**

The method based on Rare Variants found Exclusively (RVE) in cases versus controls was described by Cohen et al. (2004) in the study of high-density lipoprotein cholesterol (HDL-C). Note that RVE is more appropriately classified as a burden test because it quantifies the variants found in either cases or controls. However, I have listed RVE as a method for analyzing selected samples because it was first described by Cohen et al. (2004) in the context of comparing variant counts in individuals in the extremes.

The RVE quantifies the difference in rare variants carried between individuals at the upper and lower 5% of the HDL-C plasma distribution to mimic a case-control design [17]. The excess of rare nonsynonymous variants in the extremes are contrasted and tested by using Fisher’s exact test. By design, Fisher’s exact test is conservative. In addition, dichotomizing the trait distribution and treating the upper and lower extremes as a case-control design is discarding information which is known to be very inefficient.

### **Likelihood-Based Approach Using Selective Sampling for QTLs**

Huang and Lin (2007) introduced a likelihood-based approach for selective-genotype designs of QTL under the premise that individuals in extremes are more likely to carry disease-causing alleles. To account for trait-dependent sampling, they provide two options to deal with the selective sampling: 1) Include the phenotype information of all individuals sampled, but only include the genotype information of those in the extremes or 2) Discard trait and genotype information for individuals who are not in the selection region  $\mathcal{C}$ . Depending on the design, the likelihood can be written in full or conditional form [37].

#### Implementation:

Let  $P(Y_i|G_i, \theta)$  denote the conditional density function, indexed by  $\theta = (\alpha, \beta, \sigma^2)$

WLOG, assume we have a single locus model:

$$Y_i = \alpha + \beta_G G_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2).$$

Let  $\gamma$  represent the multi-site genotype frequencies and let  $P(G; \gamma)$  denote the density for the genotypes.

### **Two general selective-genotyping designs options**

1. To account for selective sampling, the full likelihood includes the selection probability of all individuals. However, for individuals whose phenotype fall beyond

the regions of interest, only their genotype is used in the selection probability:

$$\prod_{i=1}^n P(Y_i|G_i; \theta) P(G_i; \gamma) \prod_{i=n+1}^N \sum_G P(Y_i|G; \theta) P(G; \gamma).$$

2. When  $n$  individuals whose phenotype are below  $c_L$  or above  $c_U$  are selected for genotyping, the conditional likelihood (using  $\theta$  and  $\gamma$ ) is written as:

$$\prod_{i=1}^n P(Y_i, G_i | Y_i \in \mathcal{C}) = \prod_{i=1}^n \frac{P(Y_i|G_i; \theta) P(G_i; \gamma)}{\sum_G P(Y_i \in \mathcal{C}|G; \theta) P(G; \gamma)}.$$

Alternatively, the likelihood in terms of  $\theta$  can be written as:

$$\prod_{i=1}^n P(Y_i, |G_i, Y_i \in \mathcal{C}) = \prod_{i=1}^n \frac{P(Y_i|G_i; \theta)}{P(Y_i \in \mathcal{C}|G_i; \theta)}.$$

To account for the selection of individuals with trait above a threshold  $U$  or below a threshold  $L$ ,  $P(Y_i \in \mathcal{C}|G_i; \theta)$  is calculated as:

$$P(Y_i \in \mathcal{C}|G_i; \theta) = 1 - \Phi\left(\frac{U - \alpha - \beta_G G_i}{\sigma}\right) + \Phi\left(\frac{L - \alpha - \beta_G G_i}{\sigma}\right).$$

The primary benefit of this likelihood-based approach is that it accounts for trait-dependent sampling. Huang and Lin (2007) report that the power increases when the thresholds become more extreme [37]. However, Lander and Botstein (1989) cautioned that using individuals in the very extremes (say at the 1% extremes versus 5%) can result in spurious results as individuals in the tails may have a different trait etiology [50].

### **Kryukov et al. (2009)**

For resequencing data, Kryukov et al. (2009) presented a strategy based on collapsing and using the extreme-trait design. For individuals who do not carry disease-causing variants, the phenotype is assumed to follow a  $N(\mu, \sigma)$  distribution. Individuals carrying one or more disease-causing variant have a mean that is shifted by the quantity  $\delta$ . The trait distribution of these individuals is represented by  $N(\mu + \delta, \sigma)$ .



To select the thresholds, Kryukov et al. (2009) ranked individuals by their trait values and proposed to take the upper and lower percentiles from this list. To test for rare variant association, they used the CMC. To compare whether the carrier frequencies under the null follow a 50%/50% ratio, they applied the chi-square test. In addition to the test of rare variant carrier frequency, Kryukov et al. (2009) also compared the carrier frequency of both rare and common variants. Rather than using the stringent  $\alpha = 5 \times 10^{-8}$  significance level suggested for GWAS, they propose using exome-wide significance based on 20,000 genes for  $\alpha = 2.5 \times 10^{-6}$  [46].

### Almost-Extreme Sampling

Li et al. (2011) introduce the notion of “almost-extreme sampling” to make inference more robust to heterogeneity in the extremes. Almost-extreme sampling requires that two additional thresholds be introduced so that the very extreme tails of the distribution are trimmed. The proposed sampling scheme is implemented using the CMC. In addition to proposing almost-extreme sampling, Li et al. (2011) also showed that implementing this design in conjunction with a two-stage designs is a cost-effective strategy [54].

#### Implementation:

Assume that the trait has a  $N(\mu, \sigma^2)$  distribution and that  $Y$  can be represented as  $Y = \mu + \varepsilon$  where  $\varepsilon \sim N(0, \sigma^2)$ . As before,  $L$  and  $U$  denote the thresholds for the  $q$ th and  $(1 - q)$ th quantiles, respectively. By using the truncated normal from Johnson & Kotz (1970), a regular extreme-trait design conditional on the genotype  $G$  can be represented as:

$$f(Y|G, Y < L \text{ or } Y > U) = \frac{N(\mu, \sigma^2)}{1 - \Phi\left(\frac{U-\mu}{\sigma}\right) + \Phi\left(\frac{L-\mu}{\sigma}\right)}.$$

To perform almost-extreme sampling, two additional thresholds are introduced such that  $L_1 < Y < L_2 < \mu$  and  $U_1 < Y < U_2 < \mu$  as seen in figure 2.5.

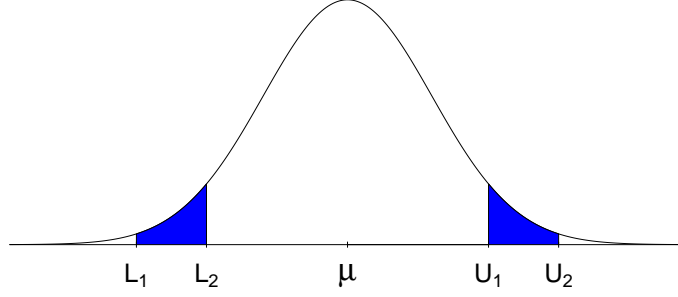


Figure 2.5: The Almost-Extreme Sampling Design.

The density is then modified to incorporate the new thresholds as [54]:

$$f(Y|G, L_1 < Y < L_2 \text{ or } U_1 < Y < U_2) = \frac{\mathbb{N}(\mu, \sigma^2)}{\Phi(\frac{L_2 - \mu}{\sigma}) - \Phi(\frac{L_1 - \mu}{\sigma}) + \Phi(\frac{U_2 - \mu}{\sigma}) - \Phi(\frac{U_1 - \mu}{\sigma})}.$$

A limitation of almost-extreme sampling is that usually the tails of a trait are enriched with individuals with causal variants.

## 2.5 Gene Discovery Approaches for Mendelian Traits

### Traditional Linkage

Genes responsible for Mendelian traits are traditionally analyzed using Linkage methods. Linkage analysis is based on the congregation of loci. There are many markers with unknown function in the genome that are known and serve as a proxy. Analysis can be carried out under a model-based or distribution-free approach. Under a model-based framework, knowledge of the disease model is required. The disease model is made up of the mode of inheritance and the variant allele frequencies in the population. For Mendelian diseases, the modes are dominant or recessive. The model-based approach relies on the recombination fraction ( $\theta$ ), the probability that two loci segregate together. Then by the LOD (i.e. Log ODDs) score method, the hypothesis of no linkage between a marker and disease loci,  $H_0 : \theta = \frac{1}{2}$ , is tested against

the alternative  $H_1 : \theta < \frac{1}{2}$  and the LOD score is based on the log of the likelihood ratio and for a single marker it is denoted as:

$$LOD(\theta) = \log_{10} \frac{L(\theta)}{L(\theta = \frac{1}{2})}.$$

The general form of  $L(\theta)$  for single-marker (i.e., two-point) linkage of Mendelian traits is  $L(\theta) = \theta^r (1 - \theta)^{n-r}$  where  $n$  is the number of offspring and  $r$  is the number of recombinants [79].

Distribution-free linkage does not require specification of the disease model and is based on the frequency of shared alleles between relatives. For alleles that are the same in a relative pair, say siblings, the alleles inherited by a common ancestor are said to be *Identical-by-Descent* (IBD) or if they are the same but inherited from different parents, alleles are said to be *Identical-by-State* (IBS). Under the null hypothesis, siblings are expected to share 0, 1, or 2 alleles with probability 1/4, 1/2, and 1/4, respectively. Under the alternative, we would observe a greater number of pairs sharing 1 or 2 alleles IBD more frequently than expected. Testing of IBD carrier frequencies can be carried out using a Chi-square test.

Linkage analysis is greatly limited by the need of family data. Many multi-generation families are required yet many Mendelian diseases are extremely rare. For complex traits, linkage-based methods perform poorly mostly because many traits do not oblige to Mendelian rules of inheritance.

## Recent Approaches

### Filtering approach

An approach to find the causal variant in Mendelian traits has been the *filtering approach*. The filtering approach, as the name implies, is based on setting criteria to reduce the list of variants from exome or whole genome sequencing to a list of candi-

date variants. This filtering approach doesn't have any formal statistical test, rather it was initially described as a "proof-of-principle" approach for identifying the causal variants in the exome. The filtering approach is based on a set of assumptions, the first is that the majority of Mendelian traits result from variants in the protein-coding regions of the gene. Other assumptions or information that can be used in filtering include using SIFT or Poly-Phen scores to predict damaging variants, removing variants found in the exomes of unaffected family members, removing variants found in public databases, etc. Ng et al. (2008) first discussed the possibility of using a filtering approach on a single exome in which they limited their analysis to non-silent variants and by successively using several filtering criteria managed to reduce a list of over 12,000 variants to ~ 1,500 variants that were more likely to be disease causing [72].

In figure 2.6, I present a diagram of the variant filtering pipeline incorporating several filtering criteria that could be considered when producing a candidate gene list as suggested by the examples presented as well as those suggested by Ku et al. (2011, 2012)[48, 47]. To demonstrate the utility of using filtering on the exome as a proof-of-principle, Ng et al. (2009) compared the exomes of four individuals with the Freeman-Sheldon syndrome and reported that the previously identified gene *MYH3* was present in all exomes [75]. The following year, Ng et al. (2009) again used filtering and reported a *de novo* missense variant in *MLL2* as the cause of Kabuki syndrome [73].

There have been filtering schemes that incorporate linkage information such as the LOD score or Identity-By-Descent (IBD) sharing. Rödelsperger et al. (2011) present a filtering algorithm based on the IBD sharing of relatives, in particular those who share both alleles IBD. In addition to reducing the candidate gene list, their method provides a quality control check to detect "sequencing errors" [90].

The filtering approach is not foolproof. One of the most common criticisms of filtering

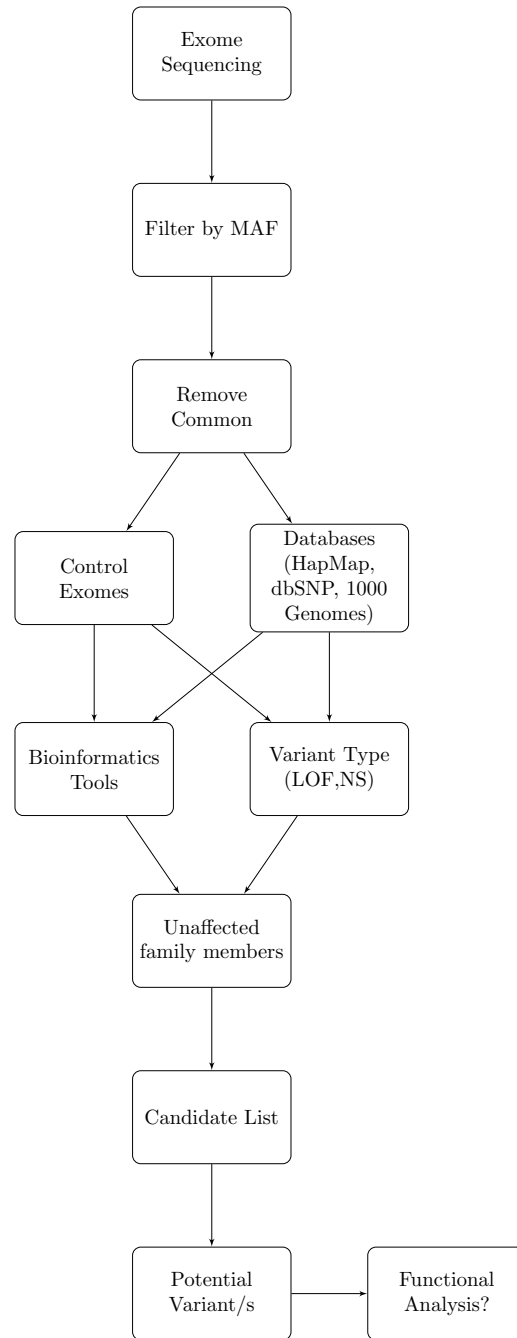


Figure 2.6: Diagram of the filtering pipeline for variant prioritization. Note that several paths are possible in filtering out variants [72, 48, 47].

is the possibility of completely missing the disease-causing variants because of genetic heterogeneity or that the causal variant went undetected in the public databases

[6]. Rödelsperger et al. (2011) also caution against using Bioinformatics tools in filtering since these often have a low sensitivity to detect causal variants, however their approach can have an increased number of false positives when variants are in linkage disequilibrium [90].

### Joint-Rank Method

The joint-rank method by Ionita-Laza et al. (2011) was designed to combine the weighed sum approach and filtering for gene discovery in Mendelian traits. This method assigns a rank to each gene based on both the weighted sum and filter-based approaches for different variant groupings. The variant groupings can be created to include only novel variants, rare variants, variants filtered out using databases, etc. The test statistic for the joint-rank method is based on tabulating the number of affected individuals with one or more alleles in the variant in question (i.e. the burden) [40].

#### Implementation:

Let  $M_A$  denote the number of novel variant positions observed in a set ( $A$ ) of sequenced affected individuals (1) at locus  $G$ . For each affected individual  $i$ , the load  $L_i$  of novel non-synonymous variants is calculated as:

$$L_i = \sum_{j=1}^{M_A} 1_{NS}(j) G_{ij}, \text{ where } 1_{NS}(i) = \begin{cases} 1 & \text{NS variant} \\ 0 & \text{otherwise.} \end{cases}$$

To rank genes, the filter-based statistic is introduced:

$$S_{filter} = \sum_{j=1}^A I_{\{L_i > 0\}}.$$

The total number of rare variants carried in affected individuals for each variant site  $j$ , denoted by  $T(j)$  is summed to form a new statistic,  $S$ :

$$S = \sum_{j=1}^M T(j).$$

$S$  can be modified to resemble a WSS-like statistic by including weights,  $w_j$ :

$$S_w = \sum_{j=1}^M w_j T(j).$$

Assuming that variant frequencies ( $f_j$ ) follow a Beta( $\alpha, \beta$ ), then  $f_j$  is estimated as:

$$\hat{f}_j = \frac{x_j + \alpha}{n_0 + \alpha + \beta}$$

where  $x_j$  is the count of times the minor allele of variant  $j$  is seen in the unaffected individuals and  $n_0$  is the total number of unaffected individuals.

The joint rank statistic is simply the average of the ranks from the weighted sum statistic ( $S_w$ ) and the filter-based statistic ( $S_{filter}$ ):

$$S = \frac{\text{Rank}(S_{filter}) + \text{Rank}(S_w)}{2}.$$

The benefit of the join-rank statistic is that it can be modified to include rare novel variants or simply the filter-based variants. In addition, this method can be applied to data based on families or unrelated individuals. A possible limitation is that the density of the WSS is approximated, which may not be appropriate in practice because of the small sample sizes expected for Mendelian trait studies.

## 2.6 Difficulties in the Quest for Rare Variants

Many of the difficulties in the detection and analysis of common variants translate to rare variants. Population substructure, or when subpopulations have different allele frequencies, can result in false positive if it goes uncorrected. For GWAS, principle components can be effective for controlling population substructure. For rare variants, the concern of population substructure is attenuated by their extremely low frequency, making population subdivision more likely. GWAS require very large sample sizes (in the hundred thousands) in order to detect common variants that have small to moderate effect sizes. While rare variants tend to have greater effect sizes,

many are *de novo* mutations and individual-specific. Recently, Nelson et al. (2012) sequenced 202 genes in 14,000 individuals and reported that nearly three quarters of rare variants were carried in one or two individuals [71].

To resolve some of the missing heritability, other types of genetic factors such as structural variants must be considered. Copy number variants (CNVs), a specific class of structural variants, have already been reported to partially contribute to the dark matter in diseases like Autism and Chron's disease [100]. In addition, Gene-Gene (GxG) and Gene-Environment (GxE) interactions may also be involved in the etiology of disease. The difficulty in analyzing such interactions goes beyond developing statistical methods; rather to make biological sense and draw plausible conclusions from them.

Recently, a shift to understand the biology of diseases has lead back to Mendelian traits. Understanding the etiology of Mendelian traits is far from simple. For example, genetic heterogeneity at the locus level disrupts the one-to-one correspondence expected in Mendelian diseases. This one-to-one correspondence can also be disrupted by imprecise disease classification. Phenotypic heterogeneity as seen in HTG and breast cancer subtypes makes robust phenotype definitions necessary. Programs such as the NIH's UDP have called for refined phenotype definitions given the advances in technology and the increased knowledge of biological mechanisms [28].

Methods implemented for rare variants are susceptible to variant misclassification. Misclassification occurs when a variant is incorrectly defined as causal or when non-causal variants are called causal and are included in the analysis. In this regard, statistical methods that are robust to non-causal variants are needed. To improve variant specification, functional analysis may be utilized to determine the variant type (e.g., non-synonymous, indel, splice-site, etc.) as well as Bioinformatics tools



(e.g., SIFT or Poly-Phen) that determine the potential of a variant to be damaging or neutral. However, many of these tools have low sensitivity and have been reported to be error-prone [26].

Current methods make use of collapsing to enrich the signal of rare variants. However, doing so makes it difficult to discern the causal variant in a grouping. Variants with opposite effects can be present at a locus, those that confer a protective effect and those that increase the susceptibility to disease. Inclusion of both types of variants can lead to a loss of power. In addition, some methods purposely exclude common variants or assume that there is no LD between rare and common variants. Yet both rare and common variants can reside on the same locus and are often jointly contribute to a common disease [18]. Ideally, methods should be robust to the inclusion of causal and non-causal variants and be able to discriminate between risk and protective variants without losing power.

## Chapter 3

# Selective Sampling in Rare Variant Association Studies

The purpose of the current chapter is to evaluate the performance of several sampling designs for rare variant-quantitative trait (QT) association studies. Practical issues related to designing efficient studies, such as threshold selection and sample size, are investigated using extensive simulations. The chapter is concluded with a discussion of the findings and implications for future sequence-based studies.

### 3.1 Introduction

Understanding of diseases with complex etiologies have been greatly enhanced over the past two decades. In particular, Genome-Wide Association Studies (GWAS) have led to the identification of many common variants (with  $MAF \leq 5\%$ ), some of which have been disease-causing. Yet, much of the heritability in complex traits remains to be explained [62] and many of the reported common variants have small effect on the trait of interest. Consider height, for example. In one study, Allen et al. (2010) reported that 180 SNPs were associated with human height but these variants collectively explain only 10% of the variance in height [1]. To explain 45% of the variance

in height, Yang et al. (2010) reported that close to 295,000 SNPs could be required [104]. Because common variants typically do not contribute significantly to the heritability of complex traits, interest has shifted to uncommon variation driven by the hypothesis that common complex traits may result from rare causative variants [31] with greater effects. Specifically, rare variants have been shown to be associated with a variety of complex traits, including colorectal adenomas, low density lipoprotein cholesterol (LDL-C) and triglyceride (TGs) levels [17, 25, 43, 91]. The identification of rare variants has been facilitated by advances in next-generation sequencing technology [94]. However, sequencing the large number of genomes at high coverage required for rare variant discovery is currently infeasible. In addition, the volume of data generated from high throughput sequencing technologies poses great challenges in both data processing and storage [84].

To reduce the sequencing cost for quantitative traits (QT), two strategies can be incorporated into the study design: sampling individuals with extreme trait values and utilizing publically available phenotyped cohorts. In order to design a cost-effective study that is adequately powered for rare variant discovery, a comprehensive evaluation of sampling strategies is imperative. To optimize power for a fixed sample size, a frequently used study design is to select individuals with extreme quantitative traits (QTs). This design has been used for genome-wide association studies [36, 51] and was used in the Dallas Heart Study to implicate rare variants in a number of metabolism-related traits [17, 91]. Extreme sampling is currently being used in exome sequencing studies including the NHLBI-Exome Sequencing Project (ESP). By studying individuals with extreme QTs, this design decreases misclassification and can enrich for causal variants. Sequencing individuals with extreme trait values has also been shown to be a cost- and time-effective approach for the mapping of complex quantitative traits [37, 51, 55].

For many complex traits, identifying causal variants associated with only one tail of the QT distribution is often of interest when the extreme QT has clinical relevance. For example, studying the genetic etiology of obesity and hypertension may be of interest. For the study of obesity where BMI is the quantitative trait of interest, often individuals with normal BMI are used as a comparison group instead of individuals with extremely low BMIs. While extremely low BMIs can be caused co-morbidities, the genetic architecture in genes of interest for BMI may be different in lean individuals than those within the clinically normal limits. For hypertension, normotensives are used as the comparison group since it is believed that those genes that modulate systolic and diastolic blood pressure in the low extremes [41] may be different from those that lead to increase these measures to a level of clinical significance.

Plomin et al. (2009) has suggested that a quantitative framework can be adapted in mapping binary or continuous complex traits [85]. Incorporating the idea of extreme-trait sampling for quantitative traits, Huang and Lin (2007) proposed a likelihood-based method that accounts for selective sampling [37] in association studies with common variants. For binary disorders, a liability threshold model can be applied where polygenic and environmental factors form a latent variable [24]. Under this model, individuals who have a liability above some threshold value exhibit the disorder while those whose liability falls below the threshold do not. In clinical practice, QTs are often dichotomized. However, it is recommended that the full QT data be used to avoid discarding information.

The purpose of this work is to determine the optimal sampling design for detecting rare variant associations under a quantitative framework. This framework is based on a general likelihood method for rare variant-complex trait association [55] that accounts for selective sampling as done for common variants by Huang and Lin (2007) [37]. By considering several factors influencing the study, the power and false positive

rate for several selective sampling schemes are evaluated using extensive simulations.

For a fixed sample size, designs that use selective sampling may result in increased power compared to population-based study designs where individuals are randomly sampled from the general population. Options to increase the pool of individuals available include using a less stringent cut-off or sampling individuals from the general population for the comparison group. Another alternative is to sample individuals from an existing public cohort where genotype and several phenotypes have been well characterized. By using the distribution quantiles to determine selection thresholds, asymmetric thresholds and unbalanced designs were compared to the symmetric balanced sampling design. In addition, the percentage/proportion of causal variants (i.e., potentially disease-causing), among all variants was varied to include the range from no variants being causal to all variants being causal. Remaining variants were neutral variants having no effect on the trait. For variants that were causal, the variant effect size (i.e., the mean shift induced by the variant) was varied and evaluated for the designs considered. Note that for neutral variants, the variant effect size would be set to zero. Finally, we explored the effect of dichotomizing the QT on power and the value of incorporating data from existing cohorts.

## 3.2 Methods

Suppose that the number of rare variants for individual  $i$  ( $i = 1, 2, \dots, n$ ) at variant site  $j$  ( $j = 1, 2, \dots, l$ ) is denoted by  $v_{ij}$  where  $v_{ij} = \{0, 1, 2\}$ . Note the  $v_{ij}$  represents the genotype of an individual, for example,  $v_{ij} = 2$  indicates that individual  $i$  is homozygous (for the minor allele) at variant site  $j$ . Similarly, a new variable  $g_{ij}$  can indicate whether an individual  $i$  carries at least one rare variant at site  $j$ , namely:

$$g_{ij} = \begin{cases} 1, & \text{if } v_{ij} = 1 \text{ or } v_{ij} = 2 \\ 0, & \text{otherwise.} \end{cases}$$

Customarily, rare variants are aggregated, say at the gene locus, since their low frequency would make it difficult to detect associations with individual variants. Thus one can group or “collapse” multi-site genotypes into  $m$  groups. We let the groups be indexed by  $k$  ( $k = 1, 2, \dots, m$ ), presumably with  $m \leq l$ . We define the collapsed genotype  $G_{ik}$ , similarly to what Basu and Pan (2011) call a “Super-Variant” [8], using the CMC collapsing scheme from Li and Leal (2008) [53] as:

$$G_{ik} = \begin{cases} 1, & \text{if } \sum_j g_{ij}^k > 0 \\ 0, & \text{otherwise.} \end{cases}$$

In words,  $G_{ik}$  serves as an indicator function of whether any rare variant is present at any of the  $j$  variant sites in group  $k$  for individual  $i$ .

Let  $Y_i$  denote the QT for individual  $i$ . For  $i \neq j$ , assume that  $\text{Corr}(Y_i, Y_j) = 0$  for individuals  $i$  and  $j$ . Then  $Y_i$  is modeled in a traditional regression framework using the collapsed genotypes  $G_{ik}$ :

$$Y_i = \beta_0 + \sum_k \beta_k G_{ik} + \varepsilon_i, \text{ with } \varepsilon \stackrel{i.i.d}{\sim} N(0, \sigma^2).$$

The  $\beta_k$  coefficients represent the effect size of the collapsed genotypes  $G_{ik}$ . Note that this representation can be easily extended to include covariates like gender, age, race, etc. In order to assess QT-rare variant association, a likelihood ratio test was chosen for the type I error and power calculations.

### 3.2.1 Sampling Framework

WLOG, the indices will be dropped for simplicity. Following Huang and Lin (2007), to implement a selective sampling scheme [37], we first let the selection region be denoted by  $S$  such that if  $y \in S$ , then the individual is sampled. For example, suppose that for a certain trait, all trait information on the population is represented by  $P = (-\infty, \infty)$  and that  $S$  represents a subset (i.e.,  $S \subseteq P$ ) of individuals whose phenotype is greater

than some threshold  $T$ , so that  $S = (T, \infty)$ . Then for an individual with phenotype  $Y$ , we define the selection indicator  $I_S$  as:

$$I_S = \begin{cases} 1, & y \in S \\ 0, & \text{otherwise.} \end{cases}$$

Since individuals are selected based on their phenotypic information, selection is conditionally independent of their genetic information, thus the probability that an individual is selected, i.e., whose phenotype falls in the selection region  $S$ , can be expressed as:

$$P(I_S = 1|Y, G) = P(I_S = 1|Y).$$

In order to account for the selective nature of the sampling scheme, the likelihood is modified by conditioning on whether an individual's phenotype falls in the selection region, and thus  $I_S = 1$ . Note that the likelihood under a random sampling scheme is not equivalent to the likelihood that accounts for selection,  $P(Y|I_S = 1, G) \neq P(Y|G)$ .

Thus the *Selection-Corrected Likelihood* is presented as:

$$P(Y|I_S = 1, G) = \frac{P(I_S = 1|Y)P(Y|G)}{\int_{y \in S} P(I_S = 1|Y)P(Y|G)dy}.$$

Depending on the approach taken to select individuals, the selection region and the selection probabilities will change. In the next section, the sampling designs implemented in this thesis are defined and the corresponding selection probabilities are expressed, as needed.

### 3.2.2 Description of Study Designs and Sample Selection

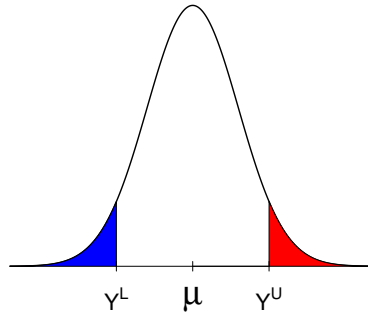
For study designs involving quantitative traits, distribution-based thresholds are usually used to define the individuals who are of clinical interest. Without loss of generality, suppose the upper extreme of the QT distribution is the tail of clinical interest. Individuals from the opposite extreme could serve as the comparison group. This type

of design is the extreme trait design or a two-tail design [105]. Individuals from the opposite extreme of the trait distribution are sometimes referred to as the “hypernormal” control group [64] or “super controls” [85]. A second design is the one-tail design where individuals with values at one extreme of the trait distribution are sampled and compared to individuals sampled from the remainder of the distribution. With a one-tail sampling design, a single threshold is specified to mark off which individuals belong to the tail of clinical interest. Another alternative is to sequence only individuals in one tail and select individuals randomly from the general population for the comparison group. The sampling designs implemented in this thesis are summarized in table 3.1 and displayed in figures 3.1(a)-(c).

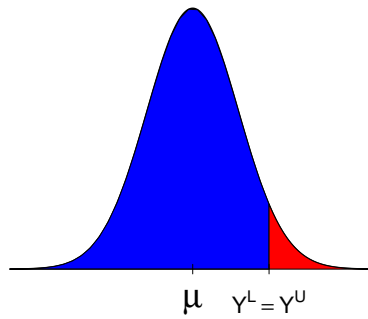
Design 1: Extreme Trait/Two-Tail (Figure 3.1(a))	Design where individuals from both extremes are sampled, i.e., two thresholds are used to mark off the upper and lower extremes of the trait distribution.
Design 2: Single Threshold/One-Tail (Figure 3.1(b))	Design in which individuals with values at one extreme of the trait distribution are sampled and compared to the individuals from the remainder of the distribution.
Design 3: Single Threshold and Random Sample (Figure 3.1(c))	One-tail sampling with the comparison group taken from the general population; In this design, a single thresholds is used to select individuals from the tail of clinical interest and the comparison group is taken to be a random sample from the general population or an existing cohort.
Balanced	A design in which the number of individuals selected from the tail of clinical interest and the comparison group are equal.

Table 3.1: Description of sampling designs.

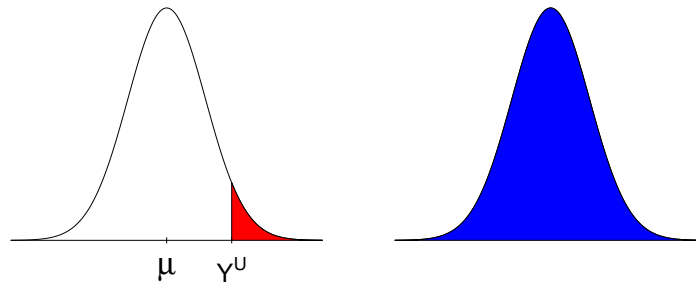




(a) Design 1: Extreme Trait Design



(b) Design 2: One-tail Design



(c) Design 3: One-Tail Design with Random Sample

Figure 3.1: Illustration of Three Sampling Designs.

Within each of these designs, balanced and unbalanced designs are explored. In the balanced design, the number of individuals sequenced from the tail of clinical interest and the comparison group are equal. In the unbalanced design, the number of individuals sequenced from each tail is not equal. We use sampling ratios such as a 1:2 ratio, where one individual from the tail of clinical interest is sampled per two individuals from the comparison group. Within each of these sampling frameworks, we further explore the effect of asymmetric thresholds on the power to detect rare variant associations. When an existing cohort is used, asymmetric thresholds may affect whether a design is balanced or not.

Returning to the representation of the selection region  $S$  and the selection probability discussed in the previous section, new notation is introduced. In order to sample individuals under a selective sampling design, say the extreme trait design (see figure 3.1(a)), a lower ( $Y^L$ ) and upper ( $Y^U$ ) threshold based on the quantiles ( $q$ ) of the trait distribution are defined as  $Y^L = \Phi^{-1}(q_L)$  and  $Y^U = \Phi^{-1}(q_U)$ , respectively. Note that for the extreme trait design, individuals with QT  $Y$  satisfying  $Y < Y^L$  or  $Y > Y^U$  are included in the sample and thus the selection region is represented as:

$$S = \{(-\infty, Y^L) \cup (Y^U, \infty)\}.$$

The selection probability in this case would be represented as:

$$P(I_S|Y = y) \propto \begin{cases} \Phi(Y^L), & Y < Y^L \\ 0, & Y^L < Y < Y^U \\ 1 - \Phi(Y^U), & Y > Y^U. \end{cases}$$

For simplicity the normalizing constant,  $\Phi(Y^L) + 1 - \Phi(Y^U)$ , is not included in the selection probability above for the extreme trait design. Note that for the one-tail design, the probability of selecting individuals for either the tail of clinical interest or the comparison group can be determined directly from the cdf, without the need of a normalizing constant, since there is only one threshold.

A symmetric extreme trait design can be implemented when the upper  $q$  and lower  $1-q$  quantiles are used to define the thresholds for selecting individuals. For example,  $Y^U = \Phi^{-1}(0.95)$  and  $Y^L = \Phi^{-1}(0.05)$  means that individuals with traits in the upper and lower 5% of the distribution will be sampled. Note that the one-tail design consists of selecting a single threshold such that  $Y^U = Y^L$ .

### 3.2.3 Approach for Analyzing Dichotomized Phenotypes

The quantitative approach is compared with the approach where the QT is analyzed as a dichotomized trait (represented by  $D$ ). In this case,  $D = 1$  if  $Y \in S$  or  $D = 0$  otherwise. For a single covariate, we can define  $\pi = P(D = 1|X)$ . When several covariates are of interest,  $\pi$  can be similarly defined and we can use logistic regression model such as:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_j \beta_j X_j + \varepsilon.$$

Alternatively, when the association can be represented by a contingency table, chi-squared or Fisher's exact test can be used. Here, Fisher's exact test is implemented.

## 3.3 Simulation Set-Up

Table 3.2 lists the information on the simulated data, which is based on African Ancestry. Details on how the genetic data was generated can be found elsewhere [55].

Gene Length (bp)	1,500 bp
Mutation Rate	$1.8 \times 10^{-8}$ /generation
Model for $s$	$\text{Gamma}(0.184, 8, 200)$

Table 3.2: Settings for simulated rare variant data based on Africans [11].

Briefly, the simulated genetic data was based the work of Boyko et al. (2008) who use specific population genetics measures such as selection and demographic events to model mutation distributions [11]. Specifically, the selection disadvantage ( $s$ ) for Africans was modeled using a Gamma distribution with parameters  $\alpha_A = 0.184$  and  $\beta_A = 8,200$ . In the current work, only NS variants are considered since these are more likely to be disease-causing or to affect protein function [87]. Furthermore, the set of causal variants is denoted by  $C$  and is integrated into the phenotype distribution as  $N(\sum_{k \in C} \beta_k G_{ik}, \sigma^2)$ . All variant sites within a gene region are analyzed but only variants with a minor allele frequency of  $\leq 0.01$  are considered to be causal. 25, 50, 75 or 100 percent of the variant sites are randomly selected to be causal and affect the QT distribution while the remaining variant sites are set to be neutral. Note that the mean is shifted by  $t$  standard deviations when the variant is causal:  $\tilde{\beta} = t\sigma$ . For simplicity, all variant groups were assumed to have the same variant effect. For the phenotypes generated, different variant effects were used, namely  $t = \{0.25, 0.5, 1.0, 1.5\}$ .

The sample sizes selected ranged between 1,000 to 10,000 individuals for each type of design. The thresholds used for the tail of clinical interest were based on the quantiles 0.90 or 0.95. For the extreme trait design, cut-offs of 0.05 and 0.10 were used. When asymmetric thresholds were used, the cut-offs for the comparison group ranged between 0.05 to 0.80 with increments of 0.05 or 0.10. For the one-tail design, the quantiles selected were 0.90 and 0.95. When sampling individuals from the general population for the comparison group, the quantile 1.00 is used to specify that the individual is not selected based on their trait value. The existing cohort was simulated with a size of  $N=10,000$  individuals. To evaluate power, 10,000 replicates were used using exome-wide significance  $\alpha = 2.5 \times 10^{-6}$ . From here on, QT denotes quantitative trait and DT denotes dichotomized trait.

## 3.4 Results

### 3.4.1 Evaluation of Type I Error

The type I errors were evaluated for the three types of designs described in table 3.1 and illustrated in figures 3.1(a)-(c). For  $\alpha = 0.05$  and  $\alpha = 0.001$ , the design parameters were varied from changing the upper and lower thresholds, using unbalanced sampling, and using asymmetric thresholds.

Design	Description	$q_L$	<u>Type I Error Rates</u>	
			DT	QT
1	Extreme-Trait Design	0.05	0.0395	0.0521
2	Single-Threshold	0.95	0.0353	0.0514
3	One-Tail and Random Sample <sup>1</sup>	1.00	0.0398	0.0505

Table 3.3: Comparison of type I error for the three designs when the upper threshold is set at  $q_U = 0.95$ , balanced sampling is implemented in which 1,000 individuals are selected from each extreme, and  $\alpha = 0.05$ .

Total Sampled	No. Upper	No. Lower	$q_L$	<u>Type I Error Rates</u>	
				DT	QT
1000	500	500	0.05	0.0345	0.0528
1500	500	1000	0.10	0.0403	0.0545
2000	500	1500	0.15	0.0362	0.0519

Table 3.4: Comparison of Type I Error for unbalanced sampling from an existing cohort with upper threshold  $q_U = 0.95$  and  $\alpha = 0.05$

Based on the simulations under several scenarios and designs, there is no significant

				<u>Type I Error Rates</u>			
				<u><math>\alpha = 0.05</math></u>		<u><math>\alpha = 0.001</math></u>	
No. Upper	No. Lower	$q_U$	$q_L$	DT	QT	DT	QT
100	900	0.99	0.09	0.0293	0.0537	0.0005	0.0012
200	800	0.98	0.08	0.0331	0.0530	0.0006	0.0011
300	700	0.97	0.07	0.0319	0.0529	0.0010	0.0016
400	600	0.96	0.06	0.0334	0.0583	0.0004	0.0010
500	500	0.95	0.05	0.0345	0.0528	0.0004	0.0018
600	400	0.94	0.04	0.0343	0.0539	0.0009	0.0015
700	300	0.93	0.03	0.0297	0.0566	0.0004	0.0012
800	200	0.92	0.02	0.0277	0.0542	0.0002	0.0010
900	100	0.91	0.01	0.0173	0.0552	0.0000	0.0014

Table 3.5: Comparison of Type I Error for  $\alpha = \{0.05, 0.001\}$  for varying thresholds and unbalanced sampling from an existing cohort of fixed size  $N = 10,000$ . A total of  $n = 1,000$  individuals selected. Because an existing cohort is used, the thresholds depend on the number selected from each tail. For example, when selecting 100 individuals from the upper tail and 900 for the lower tail,  $q_U = 1 - \frac{100}{10,000} = 0.99$  and  $q_L = \frac{900}{10,000} = 0.09$ , respectively.

discrepancies in the type I errors. When the QT is dichotomized, the type I errors are below the specified significance level  $\alpha$ . For example in the case when balanced designs are implemented for  $\alpha = 0.05$ , the type I errors are between 0.035 and 0.04 for the dichotomized trait. This is because Fisher's exact test was used and this approach is known to be conservative.

### 3.4.2 Effect of $\beta$ and Percent Causal Variants on Power

Figure 3.2 presents the power analysis for the three designs described in table 1 with varying values of  $\beta$  and percent causal variants. Two thousand individuals were

selected for sequencing from the general population, 1,000 from each extreme. The upper threshold was based on the theoretical distribution at the 95th percentile,  $Y^U = \Phi^{-1}(0.95)$ .

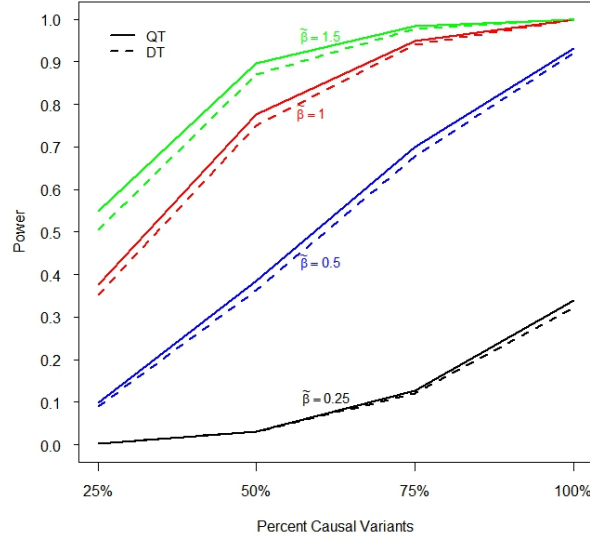


Figure 3.2: Power Comparison for Balanced Extreme Trait Design (Design 1), i.e., 1000 individuals from upper and 1000 from lower tails.  $q_U = 95\%$  and  $q_L = 5\%$ ,  $\tilde{\beta} = \{0.25, 0.5, 1.0, 1.5\}$ .

When the variant effect size is  $\beta=0.25$ , design 2 and design 3 have negligible power over the entire range of percent causal variants while for design 1 there is a modest increase in power when at least 50% of the variants are causal. For an effect size  $\beta=0.5$  and greater, the power for design 2 significantly increases as the percent of causal variants increases. Design 1 offers a power advantage over design 2 when the variant effect is  $\beta=0.5$ , but when  $\beta=\{1.0, 1.5\}$  the power differences are minimal. For example, when 25% of the variants are causal and  $\beta = 1.5$ , the power for design 1 is 0.5495 and for design 2 it is 0.5447. Design 1 is robust to dichotomizing for different values of  $\beta$  and percent causal variants. Designs 2 and 3 are sensitive to

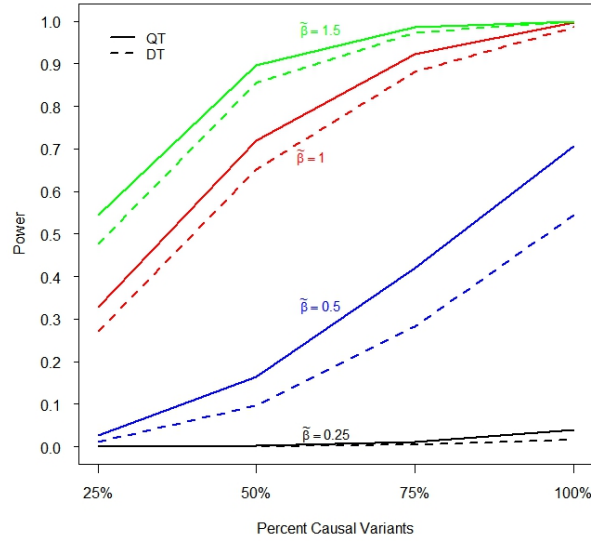


Figure 3.3: Power Comparison for Balanced One-tail (Design 2), i.e., 1000 individuals from upper and 1000 from lower tails.  $q_U = q_L = 5\%$ ,  $\tilde{\beta} = \{0.25, 0.5, 1.0, 1.5\}$ .

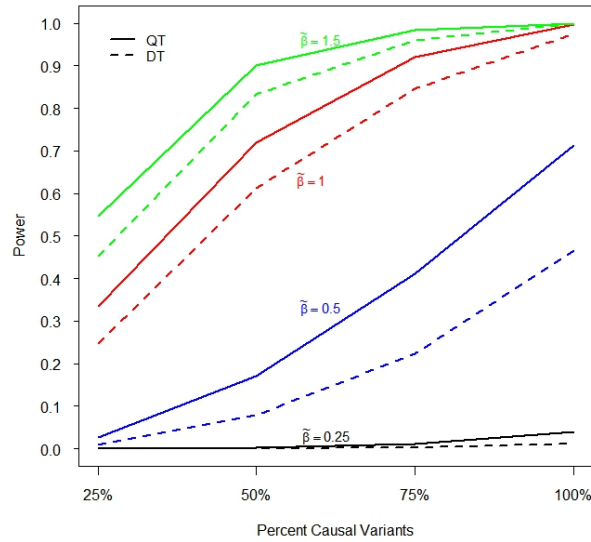


Figure 3.4: Power comparison of one-tail design with a random sample (Design 3), with 1000 individuals from upper tail at  $q = 95\%$  and 1000 individuals from the general population,  $\tilde{\beta} = \{0.25, 0.5, 1.0, 1.5\}$ .



dichotomizing and the departure between the power of the QT and dichotomized trait begins to increase as the percent of causal variants increases (see figure 3.1).

### 3.4.3 Effect of Sample Size and Threshold on Power

When an existing cohort is available, it is possible to implement an extreme trait design under a varied range of values for  $\beta$  and percent causal variants. In panel (a) of figure 3.5, 500 individuals were selected from each extreme with thresholds  $Y^U = \Phi^{-1}(0.95)$  and  $Y^L = \Phi^{-1}(0.05)$ . In panel (b), 1,000 individuals from each extreme were selected for sequencing using upper and lower thresholds set to  $Y^U = \Phi^{-1}(0.90)$  and  $Y^L = \Phi^{-1}(0.10)$ .

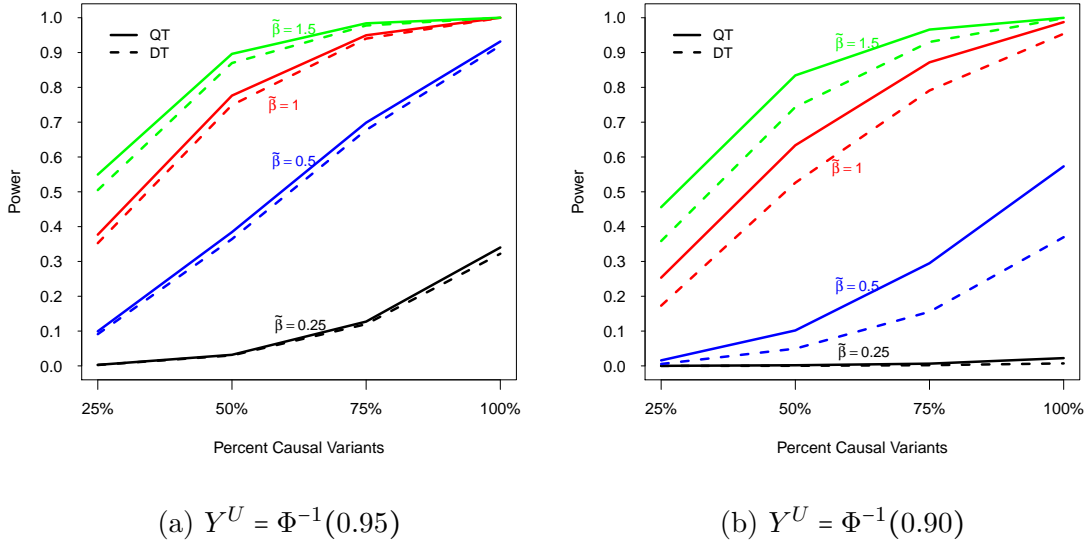


Figure 3.5: Power Plots for an Existing Cohort for (a)  $\Phi^{-1}(0.95)$  and (b)  $\Phi^{-1}(0.90)$ .

A less stringent threshold with more individuals selected leads an increase in power. For example, when  $\beta=0.5$  and 75% of the variants are causal, the power for analyzing the trait as quantitative is 0.3958 when the thresholds are  $Y^U = \Phi^{-1}(0.95)$  and  $Y^L = \Phi^{-1}(0.05)$  and 0.5818 for thresholds  $Y^U = \Phi^{-1}(0.90)$  and  $Y^L = \Phi^{-1}(0.10)$ .

As opposed to the scenario when a fixed number of individuals are to be sequenced, say 1,000 individuals from an existing cohort, in figure 3.6 I explore the gain in power if the number of individuals sequenced from the cohort increases. The percent of causal variants was set to 50%.

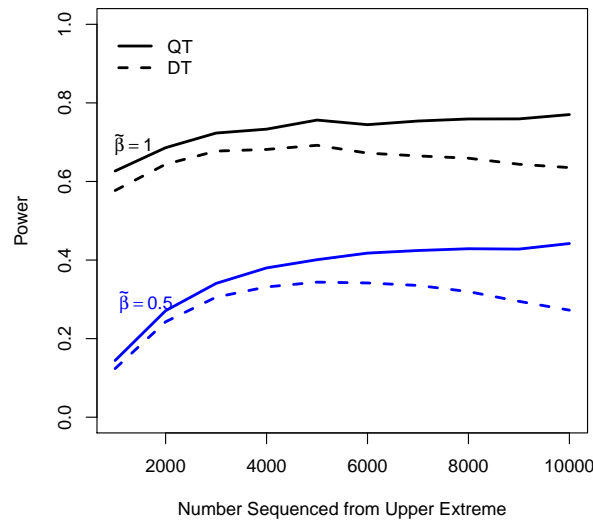


Figure 3.6: Power Plots for an Existing Cohort with 50% Causal Variants for  $\tilde{\beta} = \{0.5, 1.0\}$  under a Balanced Design.

In the case where a balanced design is used, sampling more than 4,000 individuals from an existing cohort of 10,000 does not lead to a significant gain of power when  $\beta=0.5$  or  $\beta=1$ . Specifically, for  $\beta = 1$ , increasing the number of individuals sequenced from 3,000 to 4,000 increases power to 0.7331 from 0.7232. As more individuals are selected from the entire cohort, the power begins to drop off for the dichotomized trait. For example, when  $\beta=0.5$  and a 5,000 individuals are sequenced, the power is 0.3438 compared to 0.3197 when 8,000 individuals are selected.

### 3.4.4 Effect of Asymmetric Thresholds and Unbalanced Sampling on Power

In figure 3.7, power is evaluated for the three designs under the sampling ratios of  $\{1:1, 1:2, 1:3, 1:4\}$  when  $\beta=0.5$  and 50% of the variants are causal. 1,000 individuals were sequenced from the upper tail using threshold  $Y^U = \Phi^{-1}(.95)$ .

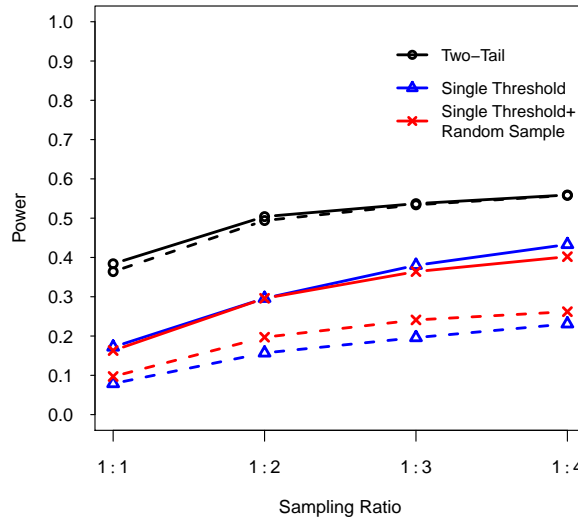


Figure 3.7: Power Comparison of Three Designs for Sampling Ratios 1:1, 1:2, 1:3, and 1:4. Power for the QT is represented by a solid line and for the DT by a dashed line.

Under the three designs, increasing the number of individuals sequenced in the comparison group from 1,000 to 2,000 (1:2 sampling ratio) leads to an increase in power. Under the 1:1 sampling ratio for the QT, the power is 0.384, 0.172 and 0.163 for designs 1, 2 and 3, respectively. Under the 1:2 sampling ratio, the power increases to 0.504, 0.296 and 0.296 for designs 1, 2 and 3, respectively. For design 1, there is a modest increase in power for the 1:3 and 1:4 sampling ratios, while for design 2 and 3 power is greater (figure 3.7). When the QT is dichotomized, there is a trivial power loss for design 1, but there is a marked drop in power for both design 2 and design

3. For example, in design 2 under a 1:3 sampling ratio the power is 0.380 for the QT and 0.196 for the dichotomized trait.

In table 3.6, sampling ratios up to 1:3 are evaluated in an existing cohort for several values of  $\beta$  and a greater range of percent causal variants. For an existing cohort of 10,000, selecting 500 individuals from each tail from the upper and lower 5 percentiles constitutes the 1:1 sampling ratio. Sequencing 1,000 (1:2) and 1,500 (1:3) individuals from the lower tail correspond to the 10th and 15th percentile of the distribution.

Sampling		Percent Causal Variants			
Ratio	$\tilde{\beta}$	25%	50%	75%	100%
1:1	0.25	0.0004/0.0004	0.0034/0.0039	0.0162/0.0195	0.0562/0.0672
	0.5	0.0183/0.0215	0.1238/0.1444	0.3591/0.3958	0.6692/0.7140
	1	0.2031/0.2348	0.5770/0.6267	0.8381/0.8682	0.9769/0.9889
1:2	0.25	0.0004/0.0005	0.0079/0.0079	0.0371/0.0393	0.1174/0.1275
	0.5	0.0417/0.0436	0.2179/0.2303	0.5117/0.5332	0.8032/0.8240
	1	0.2834/0.3022	0.6691/0.6901	0.8917/0.9069	0.9955/0.9974
1:3	0.25	0.0011/0.0014	0.0097/0.0103	0.0445/0.0481	0.1404/0.1510
	0.5	0.0536/0.0555	0.2430/0.2569	0.5437/0.5646	0.8371/0.8574
	1	0.3171/0.3252	0.6925/0.7083	0.9149/0.9250	0.9977/0.9990

Table 3.6: Table of Effect of  $\tilde{\beta}$  and percent causal variants in an unbalanced design (based on sampling ratios) when sampling from an existing cohort. The upper threshold was set to  $Y^U = \Phi^{-1}(0.95)$ . For the 1:1 sampling ratio, 500 individuals sequenced from upper extreme and 500 from the lower where  $Y^L = \Phi^{-1}(0.05)$ . Similarly for the 1:2 and 1:3 sampling ratios, 1,000 and 1,500 were sequenced from the lower threshold of  $Y^L = \Phi^{-1}(0.10)$  and  $Y^L = \Phi^{-1}(0.15)$ , respectively. Power reported as DT/QT.

For  $\beta=0.25$ , increasing the sampling ratio leads to a modest increase in power even if 100% of the variants are causal. When  $\beta = \{0.5, 1\}$ , there is a minor gain in power

when the sampling ratio is 1:2. The power gain between the 1:2 and 1:3 sampling designs is modest. For example, when  $\beta=1$  with 50% causal variants for the QT, the power is 0.6901 and 0.7083 for the 1:2 and 1:3 sampling ratios, respectively. Increasing the sampling ratio from 1:1 to 1:2 has a greater impact on power for an effect size of  $\beta=0.5$  and when at least 75% of the variants are causal. When  $\beta=1$  and 100% of the variants are causal, a 1:1 sampling ratio is sufficient to achieve power greater than 0.97.

In figure 3.8, power is evaluated when 1,000 individuals are available for sequencing from an existing cohort where the thresholds are not necessarily symmetric nor the design balanced. The interest here is to note whether power increases as the number individuals sampled from upper extreme increases.

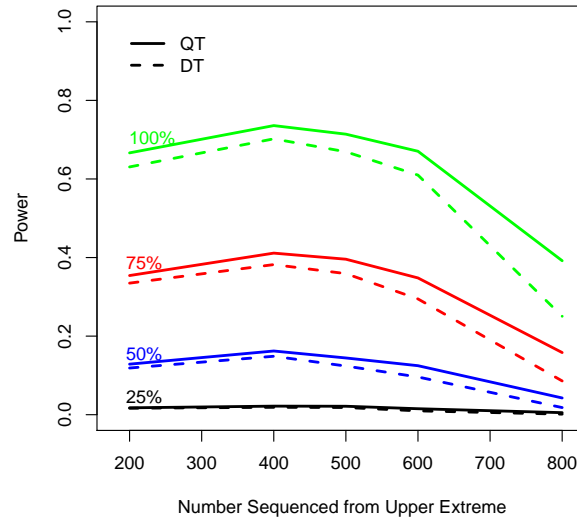


Figure 3.8: Evaluation of the effects of asymmetric thresholds and unbalanced sampling on power for a fixed sample of 1,000 from an existing cohort. The percentage of causal variants ranges between 25% to 100%.

When the percent of causal variants is greater than 50%, power is maximum when 400 individuals are sequenced from the upper extreme  $Y^U = \Phi^{-1}(0.96)$  and 600 from the lower extreme  $Y^L = \Phi^{-1}(0.06)$ . Specifically when 75% of the variants set as causal, power is 0.4113 for the QT and 0.3821 for the dichotomized trait. It is interesting to note that power drops off to 0.3481 for the QT and 0.2947 for the dichotomized trait when these numbers are reversed to 600 from the upper  $Y^U = \Phi^{-1}(0.94)$  and 400 from the lower  $Y^L = \Phi^{-1}(0.04)$  (please see figure 3.8).

To increase the pool of individuals available for the comparison group, asymmetric thresholds were considered. Power is presented in table 3.7 for a balanced design of 2,000 individuals for  $\beta=\{0.25, 0.5, 1.0\}$  and 50% causal variants with the upper threshold set to  $Y^U = \Phi^{-1}(0.90)$ . For  $\beta=1$ , the quantile for the lower bound does not affect the power significantly for the QT. For example, power is 0.7003 for the symmetric threshold  $Y^L = \Phi^{-1}(0.10)$  and 0.6708 for the threshold  $Y^L = \Phi^{-1}(0.50)$ . However, for the dichotomized trait, the power decreases steadily. When  $\beta=0.5$ , the power drops of faster for both the QT and dichotomized trait for quantiles  $q_L = 0.30$ .

$q_L$	$\tilde{\beta}$		
	0.25	0.5	1
0.1	0.0130/0.0150	0.2430/0.2712	0.6535/0.7003
0.2	0.0075/0.0093	0.2077/0.2386	0.6337/0.6864
0.3	0.0049/0.0069	0.1746/0.2074	0.6345/0.6870
0.4	0.0046/0.0054	0.1477/0.1849	0.6097/0.6714
0.5	0.0022/0.0029	0.1230/0.1624	0.6090/0.6708
0.6	0.0014/0.0019	0.1099/0.1551	0.6012/0.6703
0.7	0.0015/0.0023	0.0885/0.1326	0.5673/0.6468
0.8	0.0006/0.0020	0.0655/0.1123	0.5547/0.6456
0.9	0.0007/0.0015	0.0495/0.1018	0.5111/0.6264

Table 3.7: Power under a fixed upper threshold of  $q_U = 0.90$  but with varied lower threshold ( $q_L$ ) and with varying  $\tilde{\beta}$ . Balanced sampling is implemented with 1,000 individuals sampled from each extreme, 50% of the variants causal. Power is reported as DT/QT.

Overall, the extreme trait design (denoted as design 1) provides the optimal power under the various scenarios considered and is robust to dichotomizing the QT. When using varied sampling ratios, going beyond a 1:2 sampling ratio does not result in a significant gain in power. Under the parameter values used, the sampling pool may be increased by loosening the thresholds without significant power loss.

### 3.5 Discussion

In the past five years there has been an avalanche of rare variant association testing methods. Yet an appropriate sampling design is a crucial first step on the road to identify rare causal variants. I have addressed several practical concerns dealing with sampling designs. These include selecting appropriate sample sizes and threshold values as well as using an existing cohort. A number of non-conventional observations were made for QT based mapping of complex traits. In evaluating the type I error for the three sampling designs discussed, all designs preserve the type I error when analyzing the QT as a dichotomized trait. This is due to the use of Fisher's exact test which is known to be conservative. However, analyzing the QT as is under all three designs results with the type I error relatively close to  $\alpha$ . Yet, it is still more advantageous to adapt a quantitative framework for the analysis of complex traits. When an extreme-trait design is carried-out from an existing cohort, the power for detecting an association is strongly affected by the rare variant frequencies in each extreme as well as the size of the cohort. Therefore, selecting an equal number of individuals from each extreme is not necessarily the most powerful study. Selecting fewer individuals from the upper extreme and more samples from the lower extreme is more powerful for identifying variants that increase mean QT values. This is because the differences in causative variant frequencies are larger when an unbalanced sampling design is implemented within an existing cohort.

As shown, it is more advantageous to adapt a quantitative framework for the anal-



ysis of complex traits. At the same time, a careful definition of the phenotype is necessary along with selecting appropriate threshold to classify individuals. For some traits, very stringent thresholds may be necessary. For example, autoantibodies to 21-hydroxylase beyond the 99th percentile is a precursor to Addison’s disease, an autoimmune disorder characterized by adrenal dysfunction [5]. Complex quantitative traits that have Mendelian subtypes may also need extremely stringent thresholds.

While I have tried to provide a comprehensive analysis of study designs, limitations must be addressed. Several values of percent causal variants were explored, but it may be possible that only a fraction of the variants will be causal. In this case, much larger sample sizes will be required. Though I considered both causal and non-causal variants, I did not take into effect the directionality of the variant (i.e. protective or risk). A true signal may be missed if both protective and risk variants are collapsed into the same group. In our simplified simulations, I did not consider the effects of population stratification which can be more problematic for rare variant association than for common variants. In addition, the use of public sequence data can be a concern because using controls from different cohorts can inherently consist of heterogeneous individuals. The study from the Wellcome Trust Case Control Consortium warns of this potential limitation as there is always the possibility that the control group may include individuals that are true cases [99]. However, population stratification can be controlled for in the analysis by the use of principal components. There may also be difficulty in replicating any association signals because of the heterogeneity in samples, especially when sampling is not carried out from the same population as in the original study. A greater concern comes from the sequencing data because if targets and alignments are very different, the type I error can increase (please see Chapter 4 for a discussion on this topic).

Designing powerful and robust study designs for the analysis of rare variants is paramount in the study of complex disorders. Many novel ideas continue to emerge on dealing with the unique aspects of designing such studies. The extreme trait de-

sign is a practical design frequently used in QT studies. Studies consistently report that the extremes are the most informative because they are more likely to be enriched with causal variants [17] and that sampling individuals in the middle of the distribution can dilute any association signal and can lead to reduced power [97]. However, individuals with very extreme trait values may have a different genetic etiology than individuals with less extreme values. Recently, Li et al. (2011) suggested the approach of trimming the extremes of the distribution and thus using an “almost-extreme sampling” design to address the heterogeneity that may be present in the tails [54]. Another option for study designs is to combine extreme trait sampling schemes with the use of family data. Shi and Rao (2011) have shown that increased rare variant enrichment can be accomplished by using linkage positive families [95]. This is especially important for traits that have significant heritability.

While the cost of whole-genome sequencing drops, two-stage study designs are also an alternative design for the discovery of rare variants. By resequencing ten genes in DNA pools followed by genotyping plausible SNPs in independent samples, Nejentsev et al. (2009) reported four independent rare variants associated with Type I Diabetes [70]. The extreme trait design can also be incorporated into a two-stage design. In the first stage, only individuals in the extremes are sequenced; in the second stage, genotyping can be carried out in individuals in the remainder of the distribution (i.e. non-extremes) [54] or an independent random sample [32].

In conclusion, the results and study designs presented here are aimed at providing important guidance for implementing efficient sequence based QT studies. While many challenges remain in the design of appropriate sampling designs, many more opportunities to understand the etiology of common diseases are on the horizon.<sup>2</sup>

---

<sup>2</sup>The work presented in this chapter is expected to be part of the following paper:

Banuelos R.C., Liu D.J., and Leal S.M. *Designing Efficient Sequence-based Genetic Studies of Quantitative Traits by Incorporating Selective Sampling and Public Cohort Data*. In Progress. [7]

## Chapter 4

# On the Application of Rare Variant Association Tests to Mendelian Traits

For nearly half of the Mendelian diseases and traits found in the Online Mendelian Inheritance in Man (OMIM) catalog, a causal gene has not been identified. Linkage analysis has been the traditional method for identifying genes responsible for Mendelian traits. Linkage analysis requires pedigree data but because Mendelian traits tend to be very rare, it is difficult to ascertain enough families. Recently, an alternative has been to perform exome sequencing on affected individuals coupled with a filtering scheme to narrow the list of potential causal variants. The filtering scheme often involves removing common variants found in public genetic variant databases (e.g., dbSNP or 1000 Genomes), variants found in the exomes of non-affected family members, or variants that Bioinformatics tools have predicted to be neutral or non-causal. However, such filtering schemes are not fool-proof. Public databases can include misclassified individuals and variants which can result in false negatives, i.e., disregarded disease-causing variants. In addition, many of the Bioinformatics tools to predict whether a variant is neutral or deleterious are error-prone and the clas-

sification results vary depending on the tool used [26]. A powerful method to map genes involved in Mendelian disease is to use rare variant association methods. These rare variant association methods should be robust to extreme genetic heterogeneity, i.e., when different genetic loci or alleles within a gene result in the same phenotype. In this work, one-sided burden-based rare variant association tests are applied to dominant and recessive traits to determine the power of these tests to detect a rare variant-Mendelian trait association. A simulation-based analysis is carried out using different proportions of missing deleterious and neutral variants and varied levels of genetic heterogeneity. It is shown that one-sided burden tests perform well for small sample sizes and in the case when there is little to no locus heterogeneity.

## 4.1 Introduction

Many Mendelian disorders fall under the umbrella of rare [orphan] diseases which are typically defined as those diseases that affect less than 200,000 individuals in the U.S. but have dire health consequences [77]. For about half of the diseases and traits that have been classified as Mendelian or potentially Mendelian in the OMIM catalog, a causal gene has not been identified. As of April 16, 2013, of the 6,956 phenotypes<sup>1</sup> described in OMIM, 3,570<sup>2</sup> have some genetic information available [65]. In 2008, the NIH Office of Rare Diseases Research launched the Undiagnosed Disease Program (UDP) to identify the cause of several rare diseases [77]. The purpose of this program is to find the genetic causes of 40 hard-to-diagnose diseases per year using cutting edge sequencing technology [63]. UDP’s primary aim is not to discover disease-causing genes for monogenic diseases, yet it emphasizes the importance that single-gene disorders can have to propel knowledge of the mechanisms of disease. While the etiology of many monogenic traits is not simple, the knowledge gained from Mendelian traits will enhance our understanding of complex diseases [2].

---

<sup>1</sup>Includes traits with a “suspected Mendelian basis” [65].

<sup>2</sup>This number only reflects autosomal genes.

### 4.1.1 On Traditional Linkage Analysis

Linkage analysis has been the classical approach for identifying genes responsible for Mendelian traits. Linkage analysis is discussed in section 2.5 in more detail. Briefly, linkage analysis consists of studying the cosegregation of genes within families to localize a genetic region. Linkage analysis can be carried out in either a model-based or model-free approach. For the model-based approach, a disease model that includes the mode of inheritance and frequency of gene variant is required. For the model-free approach, a disease model need not be specified but information on the frequency of the loci under consideration is needed. Because linkage analysis often requires many pedigrees and many Mendelian traits are very rare, it is difficult to obtain or observe sufficient families that have the trait of interest. In addition, for diseases that occur sporadically or from *de novo* mutations, linkage analysis will fail. In these cases, the linkage analysis step is eliminated and one or more affected individuals undergo whole genome or exome sequencing.

### 4.1.2 On the use of NGS

Many of the major genetic discoveries in the past decade have been possible because of the advances in NGS. In special cases, it is now possible to sequence the genome or exome of a single individual to localize the disease-causing gene. Because over 85% of disease-causing variants are in the protein-coding regions of the genome [16], exome sequencing often suffices. In comparison to the number of individuals to be sequenced for a linkage analysis, exome sequencing has made it possible to identify the causal variants of several Mendelian disorders using a modest number of individuals. For example, by sequencing the exomes of ten individuals, Ng et al. (2010) identified a *de novo* mutation in the *MLL2* gene as the cause for Kabuki syndrome [73]. Since then, there have been over 100 disease-causing genes identified for Mendelian diseases [88]. Exome sequencing has also been successful in identifying genes responsible of complex traits like schizophrenia and autism spectrum disorders (ASD) [103, 78].

### 4.1.3 Exome Sequencing Coupled with Filtering Schemes

When pedigrees are available, a strategy to analyze Mendelian traits can involve performing linkage analysis followed by targeted or exome sequencing. Yet another alternative is to use some type of filtering mechanism as described in chapter 2. In this case, exome sequence data is filtered against public genetic variant databases such as dbSNP or 1000 Genomes to remove common variants. However, the filtering strategy based on existing databases is limited because these databases are not free of diseased individuals and causal variants with reduced penetrance may be observed in healthy individuals. Yet family data and Bioinformatics tools such as PolyPhen or SIFT can also be incorporated into the filtering scheme. Family data is useful because the exomes of affected and unaffected family members can be compared to search for a potentially causal variant. Bioinformatics tools can be used for variant functionality prediction or to determine the potential that a variant is damaging. The drawback of filtering based on the functionality prediction tools is that such tools can have a low specificity and a large margin of error [26].

To circumvent the problems of traditional filtering schemes, Rödelberger et al. (2011) proposed a filtering algorithm based on the Identity-By-Descent (IBD) sharing of relatives, in particular those who share both alleles IBD. In addition to reducing the candidate gene list, this method provides a quality control check to detect sequencing errors. While Rödelberger et al. (2011) attempt to counter the low sensitivity of Bioinformatics tools, their method can have an increased number of false positives when variants are in linkage disequilibrium [90]. Filtering schemes in general could be regarded as “Proof-of-principle” approaches [72] given that until recently, there had not been statistical methods designed specifically for rare variant-Mendelian trait association [40].

## 4.2 Methods

A powerful method to map genes involved in familial forms of disease is to use rare variant association methods in conjunction with exome sequencing and filtering schemes. The purpose of this work is to analyze several rare variant association tests that were originally developed for complex traits and apply them to Mendelian traits. Specifically, this research evaluates the power and type I errors of one-sided burden-type tests using an extensive simulation approach. Here “burden” does not refer to the traditional population genetics definition of “genetic burden” or “mutational load” in which selection and fitness effects are considered. “Burden” here quantifies rare genetic variant counts in affected individuals in comparison to unaffected individuals. To evaluate the robustness of these tests, realistic levels of genetic heterogeneity both at the locus and allelic level are considered as well as the effects of missing neutral and causal variants. As an approach to increase the power to identify the rare causal genetic variant, I also consider unbalanced sampling where the sampling ratio of affected to unaffected is increased.

### 4.2.1 Filtering Implementation

Prior to the association testing, I implemented a filtering scheme based on the pipeline presented in chapter 2. Given that the majority of Mendelian diseases are caused by rare mutations in the protein-coding regions of genes, I remove common variants with a  $MAF > 0.01$  and any variants found in existing variant databases. In addition, I only include nonsynonymous (NS) rare variants as these are more likely to contribute to disease. The filtering steps are shown in the following flowchart (figure 4.1):

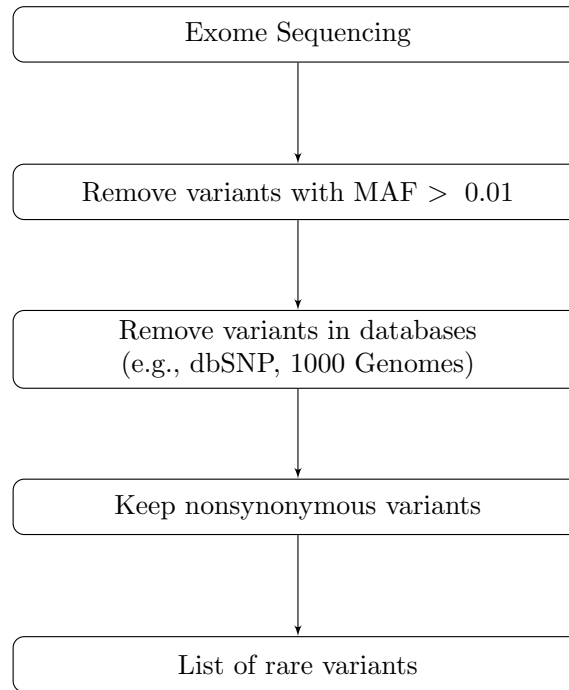


Figure 4.1: Filtering Scheme of Variants for Mendelian Traits [72, 48, 47].

### 4.2.2 Burden Tests

Type I error and power is evaluated for six commonly used burden-based rare variant association tests (listed in table 4.1 below and described in detail in §2.3.1). A burden-based rare variant association tests can be broadly defined as a test that considers the number (either in count or frequency) of rare variants in affected individuals in comparison to unaffected individuals [69]. All burden tests were implemented as one-sided since the interest is in disease-causing genetic variants. However, in section 2.3.2, I discuss the predominant tests used for bi-directional testing when there is an interest in joint association testing for protective and deleterious variants.



Burden Test	Reference
Aggregated Number of Rare Variants (ANRV)	Morris & Zeggini (2010) [67]
Combined Multivariate & Collapsing (CMC)	Li & Leal (2008) [53]
Kernel-Based Adaptive Clustering (KBAC)	Liu & Leal (2010) [56]
Rare Variants Exclusively in Cases (RVE)	Cohen et al. (2004) [17]
Variable Threshold (VT)	Price et al. (2010) [86]
Weighted Sum Statistic (WSS)	Madsen & Browning (2009) [59]

Table 4.1: Six burden-based rare variant association tests implemented for Mendelian Traits.

## 4.3 Simulation Set-up

### 4.3.1 Genetic Data

The genetic data was generated using the *Simulator of Rare Variants* (srv) script from the simuPOP software [83]. To simulate the genetic data, the demographic model for Europeans as described by Boyko et al. (2008) was used. Boyko et al. (2008) describe the distribution of the selection disadvantage coefficient  $s$  for Europeans as  $Gamma(0.206, 0.320)$  [11].

Gene Lengths (bp)	1,500, 5,000, 10,000
Variant Sites	Range of 18-50/gene
Mutation Rate	$1.8 \times 10^{-8}$ /generation
Model for selection coefficient ( $s$ )	$Gamma(0.206, 0.320)$

Table 4.2: Settings for simulated data based on Europeans [11, 83]

The purpose of the selection coefficients for variant sites is to differentiate synonymous variants from nonsynonymous variants. Fitness was modeled under a multiplicative

model and “a finite-sites mutation model with mutation rate of  $1.8 \times 10^{-8}$ ” for variants sites [83]. The gene length considered was 1,500 basepairs (bp). The number of variant sites per gene simulated selected randomly and was anywhere between 18 to 50 sites per gene.

To mimic errors in the sequencing and genotyping process such as sites not being captured during genotyping, a fraction of causal/deleterious and neutral sites were set to missing. Of the two, I expect that the proportion of deleterious sites that are missing will have a greater impact on the power to detect an association. In addition, the fraction of causal variants and the proportion of locus heterogeneity each ranged between 0 to 100%. Finally, allelic heterogeneity was treated as either present (True) or not (False).

Proportion of Missing Causal Variants	0 to 10%
Proportion of Missing Neutral Variants	0 to 1%
Proportion of Causal Variants	0 to 100%
Proportion of Locus Heterogeneity	0 to 100%

Table 4.3: Specification of noise introduced in the simulation.

### 4.3.2 Scenarios Considered

Table 4.4 provides a list of the scenarios considered and the simulation set-up of this research. I consider both dominant and recessive modes of inheritance in this work. However, I report the results for a dominant trait and note otherwise or when there is a significant difference in the results for a dominant versus a recessive trait. I evaluate the type I error and power for sampling an equal number of affected (i.e., cases) and unaffected (i.e., controls) individuals and compare these results to those of varying

sampling ratios. The sampling ratios considered range from 1:1 (corresponds to an equal number of cases and controls) up to the 1:5 ratio which corresponds to sampling 5 controls per affected individual.

Disease Models	Dominant, Recessive
Number of Cases	5, 10, 15, 20-50
Number of Controls	5*, 10, 15, 20-50
Sampling Ratios	1:1 upto 1:5

\*all tests except RVE

Table 4.4: Description of sample sizes considered, as well as scenarios implemented in the simulation study.

## 4.4 Results

Since nonsynonymous (NS) variants are more likely to be causal, only NS variants were considered in the analysis, as shown in the filtering flowchart (figure 4.1). For all simulations, 2,000 replications were carried out under significance of 0.05, 0.01, and 0.001. For permutation-based tests, 2,000 permutations were performed with an adaptive threshold set at 500 once significance was within  $10^{-6}$  of  $\alpha$ .

### 4.4.1 Evaluation of Type I Error

To evaluate the type I error, the proportion of individuals who are affected (i.e. “cases”) as a result of the disease locus was set to 0%. In addition, allelic heterogeneity was set to present (True). To circumvent that the simulation program used did not allow for the proportion of causal variants to be set to 0%, this proportion was set to  $1 \times 10^{-10}$ . No additional noise was introduced, i.e., both proportion of missing causal and proportion missing neutral variants was set to 0%.

### Equal Number of Cases and Controls

Table 4.5 presents the type I error when an equal number affected and unaffected individuals are sampled for a dominant trait when  $\alpha = 0.05$ . All tests are conservative in that the type I error remained well below  $\alpha$  with WSS being the least conservative as the sample increases.

Test	Number of Case/Controls				
	10/10	20/20	30/30	40/40	50/50
ANRV	0.0005	0.0010	0.0035	0.0025	0.0055
CMC	0.0005	0.0020	0.0090	0.0120	0.0190
KBAC	0.0005	0.0010	0.0055	0.0035	0.0065
RVE	0.0005	0.0030	0.0105	0.0120	0.0200
VT	0.0005	0.0015	0.0030	0.0020	0.0050
WSS	0.0025	0.0145	0.0290	0.0400	0.0405

Table 4.5: Type I error comparison for a dominant trait and an equal number of cases and controls,  $\alpha = 0.05$

When  $\alpha = 0.001$  or  $\alpha = 0.01$ , the type I error rates of the six test were similar to those in table 4.5.

### Unequal Number of Cases and Controls

Table 4.6 presents the results for  $\alpha = 0.05$  but with a the number of affected individuals fixed to 10 and allowing the number of unaffected individuals vary between 10 to 50. Observe that the type I error has increased and is affected by the unbalanced sampling. The type I error of WSS is drastically inflated beyond 0.05 when 30 or more unaffected individuals are sampled. For the remaining of the tests, the type I error is still controlled but more of  $\alpha$  is spent than in the case when an equal number

of individuals was sampled. For the recessive trait, all tests except for WSS produce the same type I errors. WSS was much more conservative for the recessive trait with all type I errors markedly below 0.05. For example, the type I error for sampling ratio 1:3 for WSS was 0.001.

Test	<u>Number of Controls</u>				
	10	20	30	40	50
ANRV	0.0005	0.0025	0.0040	0.015	0.0080
CMC	0.0005	0.0025	0.0145	0.0160	0.0165
KBAC	0.0005	0.0025	0.0050	0.0170	0.0090
RVE	0.0005	0.0025	0.0135	0.015	0.0125
VT	0.0005	0.0025	0.0035	0.0155	0.0080
WSS	0.0025	0.0165	0.0985	0.0825	0.0755

Table 4.6: Type I error comparison for a dominant trait under unbalanced sampling, i.e., for 10 cases and  $\{10, 20, 30, 40, 50\}$  controls,  $\alpha = 0.05$ .

In table 4.7, the same unbalanced design of 10 cases to varying number of control individuals was considered but with  $\alpha = 0.01$ . A similar trend is observed as was the case in previous tables, all tests preserve the type I error with WSS being the least conservative.

Similarly, when the exact unbalanced scenarios as in tables 4.6 and 4.7 are considered under a significance level of 0.001, all tests except for WSS perform conservatively with the type I error markedly below 0.001. For WSS, the type I error is maintained up to the 1:3 sampling ratio, but increases to 0.0015 and 0.004 for ratios 1:4 and 1:5, respectively. While all tests are conservative as in the previous tables, the tests are all more strictly conservative by not spending much of  $\alpha$  even as the number of

Test	<u>Number of Controls</u>				
	10	20	30	40	50
ANRV	0.0000	0.0005	0.0005	0.0015	0.0010
CMC	0.0000	0.0005	0.0020	0.0020	0.0015
KBAC	0.0000	0.0005	0.0005	0.0025	0.0010
RVE	0.0000	0.0005	0.0020	0.0020	0.0015
VT	0.0000	0.0005	0.0010	0.0015	0.0005
WSS	0.0000	0.0015	0.0120	0.0145	0.0220

Table 4.7: Type I error comparison for a dominant trait under unbalanced sampling, i.e. for 10 cases versus {10, 20, 30, 40, or 50} controls,  $\alpha = 0.01$ .

unaffected individuals increases.

#### 4.4.2 Power Analysis

For all scenarios in which the interest is to note the effect of genetic heterogeneity or noise on the power, the percentage of causal variants was set to 10% unless otherwise specified.

#### Exploratory Analysis of Noise Effects using CMC

In order to assess the impact of allelic heterogeneity and missing causal and neutral variants on the power to detect a rare variant association, the CMC is used because it is simple and provides analytic power. The simulation is set up for 10 cases and 10 controls with 10% causal variants, 25% locus heterogeneity and allelic heterogeneity either present or not present and significance of  $\alpha = 0.001$ . The first line of table 4.8 is the case of no missing variants (0%). When no there is no allelic heterogeneity power is 0.6525 but in the presence of allelic heterogeneity the power drops slightly to 0.6440. Lines 2 and 3 present the case of 1% missing neutral variants and 1% missing

causal variants, respectively. It can be noted that the 1% missing neutral variants has almost no effect on the power while the 1% missing causal variants causes the power to drop slightly. In the case of having both 1% missing neutral variants and 1% missing causal variants (line 4), the change in power is not much different than the case when 1% of the causal variants are missing (line 3).

Case	% Missing	<u>Allelic Heterogeneity</u>	
		Present	Not Present
1	0%	0.6440	0.6525
2	1% Neutral	0.6520	0.6505
3	1% Causal	0.6355	0.6415
4	1% Neutral, 1% Causal	0.6360	0.6425

Table 4.8: Effects of missing causal and/or neutral variants on the power of the CMC for 10 cases and 10 controls for 25% locus heterogeneity and 10% causal variants,  $\alpha = 0.001$ .

Under the similar settings, but having no allelic heterogeneity present and only increasing the percentage of missing causal variants, the power drops to 0.6265 and 0.5740 for 2% and 5% missing causal variants, respectively. However, when both the percentage of missing causal and missing neutral variants increases to 10% each, the power of the CMC drops to 0.4950, which is most likely driven by the large percentage of missing causal variants. In the case when allelic heterogeneity is present and there is 5% missing causal variants and no missing neutral variants, power is 0.5640 which is not much of a difference in comparison to not having allelic heterogeneity.

In the case that common variants are not excluded and there is no missing variants, the power is drastically reduced with allelic heterogeneity present or not. Specifically, keeping all settings the same as in table 4.8, but not excluding the common variants,

the power drops to 0.4190 and 0.4205 for allelic heterogeneity present and not present, respectively.

### **Equal Number of Cases and Controls**

For the case when only five affected individuals are available, I use only the ANRV, CMC, KBAC, RVE, and VT tests. I first consider the case in which five controls are available. When no causal or neutral variants are missing and there is between 0% to 75% locus heterogeneity, both the CMC and the RVE have zero power to detect association. Under the same scenario, ANRV, KBAC, and VT had power no greater than 0.085.

Figure 4.2 presents power as the number of affected individuals (i.e., cases) increases. The same number of unaffected individuals are selected. The settings for this scenario are: allelic heterogeneity is present, 75% locus heterogeneity, and 1% missing neutral and 1% missing causal variants. The six tests all have a similar trend, with the KBAC having superior power over the other five tests and WSS having the lowest.

Table 4.9 presents the power to detect an association for 10 affected and 10 unaffected individuals when there is 10% causal variants, 10% of causal variants are missing (i.e., 10% of the 10% causal variants are missing), and 10% of neutral variants are set to missing. The interest here is to note the change in power as the proportion of locus heterogeneity increases. The significance level was set to 0.001. As one case see, when the proportion of locus heterogeneity is set to 0 (i.e, a disease results from a single locus), all tests attain the greatest power. However, setting the proportion of locus heterogeneity to 0.25 reduces the power for most tests almost by half and further increasing the locus heterogeneity makes the power approach zero.

Keeping all other parameters the same, but fixing the proportion of locus heterogene-



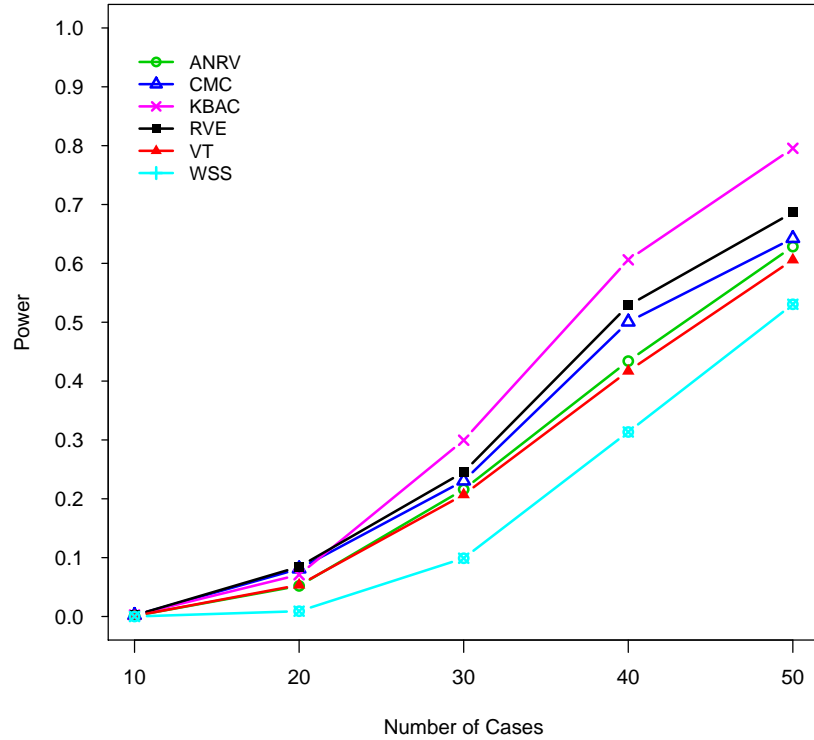


Figure 4.2: Power versus number of cases for the six burden tests. An equal number of affected (cases) and unaffected individuals are selected for a dominant trait. Settings are: locus heterogeneity fixed to 75%, allelic heterogeneity present, and  $\alpha = 0.001$ .

ity to 50% and allowing the percentage of causal variants to vary has a detrimental effect on power as seen in table 4.10. However, this drop in power is due to the high locus heterogeneity. What this low power indicates is that high locus heterogeneity will dilute an association even when the majority of variants under consideration are causal.

	<u>Proportion of Locus Heterogeneity</u>			
	0.00	0.25	0.50	0.75
ANRV	0.8805	0.3635	0.0545	0.0020
CMC	0.9140	0.4950	0.0930	0.0025
KBAC	0.9450	0.4290	0.0630	0.0015
RVE	0.9175	0.5005	0.0935	0.0025
VT	0.8805	0.3635	0.0515	0.0005
WSS	0.9065	0.4445	0.0715	0.0010

Table 4.9: Power comparison of the six burden tests for varied levels of locus Heterogeneity, fixed proportion of causal variants of 10% for a dominant trait, 10 cases versus 10 controls,  $\alpha = 0.001$ .

	<u>Proportion of Causal Variants</u>			
	0.10	0.25	0.50	0.75
ANRV	0.0545	0.0595	0.0715	0.0595
CMC	0.0930	0.0950	0.0920	0.0975
KBAC	0.0630	0.0660	0.0840	0.0705
RVE	0.0935	0.0955	0.0950	0.0975
VT	0.0515	0.0525	0.0735	0.0560
WSS	0.0715	0.0700	0.0690	0.0735

Table 4.10: Power comparison of the six burden tests, assuming a dominant trait, with varied levels of the proportion/percentage of causal variants. The locus heterogeneity was set to 50%, with 10 cases versus 10 controls, and  $\alpha = 0.001$ .

### Unequal Number of Cases and Controls

Figure 4.3 compares the power of sampling an equal number of cases and controls to the scenario of unequal sampling. All tests of association for 5 affected versus 5 unaffected were significantly underpowered and thus not included in figure 4.3. In the case when there is no allelic heterogeneity and a low proportion of locus heterogeneity (i.e., 20% or less) doubling to 10 controls and tripling (to 15 controls), the number of controls leads to gains in power for the five tests used. However, this simulation was based on having no missing causal or neutral variants, thus, it may be less realistic. But again, in the previous sections, the evaluation of the type I error and power revealed that locus heterogeneity had greatly affected power regardless of the number of causal variants present.

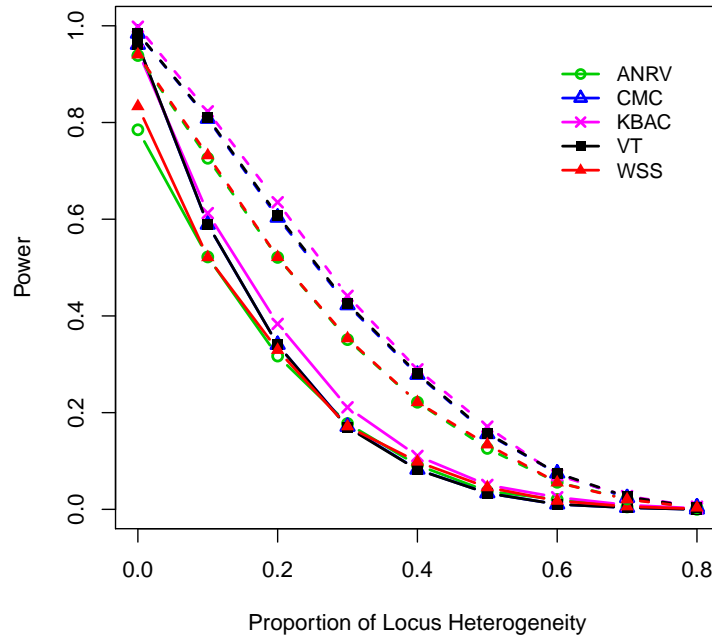


Figure 4.3: Plot of power versus proportion of locus heterogeneity for a dominant trait, 5 cases and 10 controls (solid line) or 15 controls (dashed line),  $\alpha = 0.001$ .

In table 4.11, the scenario of having 10 affected and 10 unaffected individuals is compared to the scenario when the number of unaffected individuals is doubled. As opposed to the previous unbalanced example, here 10% of the causal variants are missing, 10% of the neutral variants are missing, and locus heterogeneity is set to 25%. In addition, the proportion of causal variants ranges between 10% to 75%. As expected, increasing the proportion of causal variants has a minimal effect on power. However, doubling the number of controls from 10 to 20 significantly increases the power of all test.

	10 Cases vs. 10 Controls				10 Cases vs. 20 Controls			
	Proportion of Causal Variants				Proportion of Causal Variants			
	0.10	0.25	0.50	0.75	0.10	0.25	0.50	0.75
ANRV	0.3635	0.3640	0.3620	0.3715	0.7290	0.7400	0.7325	0.7440
CMC	0.4950	0.4975	0.4865	0.4975	0.7825	0.7910	0.7940	0.8005
KBAC	0.4290	0.4345	0.4275	0.4355	0.8285	0.8350	0.8465	0.8410
RVE	0.5005	0.5025	0.4905	0.5010	0.7860	0.7945	0.7980	0.8035
VT	0.3635	0.3640	0.3620	0.3715	0.7520	0.7580	0.7560	0.7645
WSS	0.4950	0.4975	0.4865	0.4975	0.8485	0.8580	0.8715	0.8725

Table 4.11: Power comparison for the six burden tests and a dominant trait for varied levels of percent/proportion of causal variants, locus heterogeneity fixed to 25%, 10 cases versus 10 or 20 Controls, and  $\alpha = 0.001$ .

## 4.5 Discussion

In conclusion, the six tests evaluated are very conservative in preserving type I error when an equal number of individuals are used for either a dominant or recessive trait. When unbalanced sampling is in place, all tests except for WSS, are increasingly conservative as the sampling ratio of affected to unaffected increases. The WSS, on the

other hand, has inflated type I error for a sampling ratio of 1:3 or greater.

It was also shown that missing causal variants can decrease power whereas the proportion of missing neutral variants does not affect power significantly. In the case when there is a large portion of causal variants and high locus heterogeneity, the power to detect an association will be diminished. Finally, power can be increased by doubling or tripling the number of unaffected individuals in comparison to affected individuals.

Mendelian diseases are supposed to serve as the model for understanding the molecular mechanisms of disease because of the expected one-to-one correspondence between a single gene and the phenotype. However, Mendelian traits are complicated by several factors. Many Mendelian traits have reduced penetrance or have significant locus heterogeneity [15]. In the case of reduced penetrance, the issue is then to identify possible modifier genes that also influence the phenotype [21]. This reduced penetrance may suggest that the distinction between Mendelian and complex disease is not much. In fact, there are several common traits that have monogenic forms such as early onset Alzheimer's [82] and hypo-/hyper-triglyceridemia [42]. If this is the case, another way to interrogate Mendelian diseases could be as the extremes of complex traits [3].

While exome sequencing has taken away the focus of the family as they primary means to localize disease-causing genes, family data is still pivotal for monogenic disease mapping. In particular, linkage peaks can be interrogated by using exome sequencing [10] or offspring exomes can be compared to parental exomes for identifying *de novo* variants [6]. The approach by Rödelberger et al. (2011) on using IBD information in filtering variants [90] has further demonstrated the value of family data for Mendelian disease.

## **What Does the Future hold for Mendelian Traits?**

Until the cost of whole-genome sequencing reaches the coveted \$1,000, exome sequencing will continue to play a significant role in Mendelian gene discovery. Besides accelerating discovery of disease-causing genes, exome sequencing has also shown it's potential as a diagnostic tool for many Mendelian diseases [48]. The renewed interest in monogenic diseases may foster advancements in gene mapping and in turn elucidate our understanding of diseases across the entire spectrum.

## Chapter 5

# Conclusion and Future Work

Designing an efficient sampling study is a crucial first step for identifying disease-causing rare variants. In chapter 3, I have addressed several practical concerns dealing with sampling designs for the study of rare variant-complex QT associations. I discussed selecting appropriate sample sizes and thresholds for selecting individuals as well as the use of existing public cohorts to increase power. A number of non-conventional observations were made. In evaluating the type I error for the three sampling designs discussed, all designs preserve the type I error when analyzing the QT as a dichotomized trait. This is because Fisher's exact test was used, which is known to be conservative. However, analyzing the QT as is, i.e., using the full QT information, under all three designs results with the type I error relatively close to  $\alpha$ . When an extreme-trait design is carried-out from an existing cohort, the power for detecting an association is strongly affected by the rare variants frequencies in each extreme as well as the size of the cohort. Therefore, selecting an equal number of individuals from each extreme is not necessarily the most powerful study. Selecting fewer individuals from the upper extreme and more samples from the lower extreme is more powerful for identifying variants that increase mean QT values. This is because the differences in causative variant frequencies are larger when an unbalanced sampling design is implemented within an existing cohort.

In chapter 4, variant filtering was implemented prior to evaluating the performance of six burden-based tests (ANRV, CMC, KBAC, RVE, VT, and WSS) for rare variant association in Mendelian traits. The six tests were very conservative in preserving type I error when an equal number of individuals were used for either a dominant or recessive trait. In unbalanced sampling, all tests except for WSS, were increasingly conservative as the sampling ratio of affected to unaffected individuals increased. The WSS, on the other hand, had inflated type I error for a sampling ratio of 1:3 or greater. It was also shown that missing causal variants decreased power whereas the proportion of missing neutral variants did not affect power significantly. When there was a large portion of causal variants and high locus heterogeneity, the power to detect an association diminished. Finally, power increased by doubling or tripling the number of unaffected individuals in comparison to affected individuals.

Because the focus of rare genetic variants as drivers of disease is relatively new, there are several directions that one can take. Throughout the previous three chapters, I alluded to potential research questions. Here, I formally discuss recent advancements and future directions that may hold promise for statistical applications. While the cost of whole-genome sequencing drops to the expected \$1,000, two-stage study designs are an alternative design for the discovery of rare variants. A two-stage design can be implemented where in the first stage whole or exome sequencing is performed on a group of individuals for gene discovery and the second stage only involves targeted resequencing or genotyping of the top hits in a different set of individuals. Two-stage designs have previously been implemented for common diseases. For example, by resequencing ten genes in DNA pools followed by genotyping a reduced number of plausible SNPs in independent samples, Nejentsev et al. reported four independent rare variants associated with Type I Diabetes [70]. For common QTs, the extreme trait design can also be incorporated into a two-stage design. In the first



stage, only individuals in the extremes can be sequenced and in the second stage, genotyping can be carried out in individuals in the remainder of the distribution (i.e. non-extremes) [54] or an independent random sample [32].

There are two major recurring concerns when rare or common variant are discovered. The first is that the variant does not explain enough of the variation in the trait, as discussed in chapter 2. The second, and the more relevant concern is that the biological mechanism of the variant is difficult to discern. For the first concern, a possible explanation is that other genetic factors and mechanisms such as structural variants (e.g., CNVs) and epigenetic effects also underlie the trait. Interactions between genetic and/or environmental components can also explain some of the variability. For the second concern, the biological mechanism will most likely require of functional analysis. A more practical approach may require a holistic approach such as pathway-based analysis in which both concerns could potentially be addressed.

# Bibliography

- [1] H. L. Allen, K. Estrada, G. Lettre, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.
- [2] S. E. Antonarakis and J. S. Beckmann. Mendelian disorders deserve more attention. *Nature Review Genetics*, 7(4):277–282, 2006.
- [3] S. E. Antonarakis, A. Chakravarti, J. C. Cohen, and J. Hardy. Mendelian disorders and multifactorial traits: The big divide or one for all? *Nature Reviews Genetics*, 11(5):380–384, 2010.
- [4] Applied Biosystems. *Sequencing Instruments*. Available at: <http://www.appliedbiosystems.com>, Accessed July 17, 2012.
- [5] P. R. Baker, E. E. Baschal, P. R. Fain, et al. Haplotype analysis discriminates genetic risk for DR3-associated endocrine autoimmunity and helps define extreme risk for addison’s disease. *Journal of Clinical Endocrinology and Metabolism*, 95(10):E263–E270, 2010.
- [6] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews Genetics*, 12(11):745–755, 2011.
- [7] R. C. Banuelos, D. J. Liu, and S. M. Leal. Designing efficient sequence-based genetic studies of quantitative traits by incorporating selective sampling and public cohort data. Unpublished paper.
- [8] S. Basu and W. Pan. Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 35(7):606–619, 2011.
- [9] G. Bhatia, V. Bansal, O. Hasimendy, N. J. Schork, E. J. Topol, K. Frazer, and V. Bafna. A covering method for detecting associations between rare variants and common phenotypes. *PLoS Computational Biology*, 6(10):e1000954, 2010.
- [10] D. W. Bowden, S. S. An, N. D. Palmer, et al. Molecular basis of a linkage peak: exome sequencing and family-based analysis identify a rare genetic variant in

- the *ADIPOQ* gene in the IRAS Family Study. *Human Molecular Genetics*, 19(20):4112–4120, 2010.
- [11] A. R. Boyko, S. H. Williamson, A. R. Indap, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genetics*, 4(5):e1000083, 2008.
  - [12] M. Cargill, D. Altshuler, J. Ireland, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, 22(3):231–238, 1999.
  - [13] F. Casals, Y. Idaghdour, J. Hussin, and P. Awadalla. Next-generation sequencing approaches for genetic mapping of complex diseases. *Journal of Neuroimmunology*, 2012.
  - [14] A. Chakravarti. Population genetics-making sense out of sequence. *Nature Genetics*, 21:56–60, 1999.
  - [15] A. Chakravarti and A. Kapoor. Mendelian puzzles. *Science*, 335(6071):930–931, 2012.
  - [16] M. Choi, U. I. Scholl, W. Ji, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*, 106(45):19096–19101, 2009.
  - [17] J. C. Cohen, R. S. Kiss, A. Pertsemlidis, Y. L. Marcel, R. McPherson, and H. H. Hobbs. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685):869–872, 2004.
  - [18] J. C. Cohen, A. Pertsemlidis, S. Fahmi, S. Esmail, G. L. Vega, S. M. Grundy, and H. H. Hobbs. Multiple rare variants in *NPC1L1* associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proceedings of the National Academy of Sciences*, 103(6):1810–1815, 2006.
  - [19] J. C. Cohen, A. Pertsemlidis, I. K. Kotowski, R. Graham, C. K. Garcia, and H. H. Hobbs. Low LDL cholesterol in individuals of african descent resulting from frequent nonsense mutations in *PCSK9*. *Nature Genetics*, 37(2):161–165, 2005.
  - [20] S. P. Dickson, K. Wang, I. Krantz, H. Hakonarson, and D. B. Goldstein. Rare variants create synthetic genome-wide associations. *PLoS Biology*, 8(1):e1000294, 2010.
  - [21] K. M. Dipple and E. R. McCabe. Modifier genes convert “simple” Mendelian disorders to complex traits. *Molecular genetics and metabolism*, 71:43–50, 2000.

- [22] D. F. Easton, K. A. Pooley, A. M. Dunning, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093, 2007.
- [23] M. J. Edmond, T. Louie, J. Emerson, et al. Exome sequencing of extreme phenotypes identifies *DCTN4* as a modifier of chronic *Pseudomonas aeruginosa* infection in cystic fibrosis. *Nature Genetics*, 2012.
- [24] D. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics*, 29(1):51–76, 1965.
- [25] N. S. Fearnhead, J. L. Wilding, B. Winney, et al. Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proceedings of the National Academy of Sciences*, 101(45):15992–15997, 2004.
- [26] S. E. Flanagan, A. M. Patch, and S. Ellard. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic Testing and Molecular Biomarkers*, 14(4):533–537, 2010.
- [27] W. A. Gahl, C. F. Boerkoel, and M. Boehm. The NIH Undiagnosed Diseases Program: bonding scientists and clinicians. *Disease Models & Mechanisms*, 5(1):3, 2012.
- [28] W. A. Gahl, T. C. Markello, C. Toro, et al. The National Institutes of Health Undiagnosed Diseases Program: Insights into rare diseases. *Genetics in Medicine*, 14(1):51–59, 2011.
- [29] Genome Reference Consortium. *Human Genome Assembly GRCh37.p8*. Available at: <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/data>, Accessed July 17, 2012.
- [30] J. Gonzalez-Bosquet and S. J. Chanock. Principles of analysis of germline genetics. In *Human Genome Epidemiology*, pages 13–35. Oxford University Press, New York, 2010.
- [31] I. P. Gorlov, O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz, and C. I. Amos. Shifting paradigm of association studies value of rare single-nucleotide polymorphisms. *American Journal of Human Genetics*, 82:100–112, 2008.
- [32] L. T. Guey, J. Kravic, O. Melander, et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genetic Epidemiology*, 35(4):236–246, 2011.
- [33] F. Han and W. Pan. A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, 70(1):42–54, 2010.

- [34] Helicos Bioscience Corporation. *HeliScope<sup>TM</sup> Single Molecule Sequencer*. Available at: <http://www.helicosbio.com>, Accessed July 16, 2012.
- [35] L. A. Hindorff, J. MacArthur, A. Wise, H. A. Junkins, P. N. Hall, A. K. Klemm, and T. A. Manolio. *A Catalog of Published Genome-Wide Association studies*. Available at: <http://www.genome.gov/gwastudies>, Accessed July 16, 2012.
- [36] A. Hinney, T. T. Nguyen, A. Scherag, et al. Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (*FTO*) variants. *PLoS One*, 2(12):e1361, 2007.
- [37] B. E. Huang and D. Y. Lin. Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics*, 80:567–576, 2007.
- [38] Ion Torrent. *PGM<sup>TM</sup> and Proton<sup>TM</sup> Sequencers*. Available at: <http://www.iontorrent.com>, Accessed July 20, 2012.
- [39] I. Ionita-Laza, J. D. Buxbaum, N. M. Laird, and C. Lange. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genetics*, 7(2):e1001289, 2011.
- [40] I. Ionita-Laza, V. Makarov, S. Yoon, B. Raby, J. Buxbaum, D. L. Nicolae, and X. Lin. Finding disease variants in Mendelian disorders by using sequence data: Methods and applications. *The American Journal of Human Genetics*, 89:701–712, 2011.
- [41] W. Ji, J. N. Foo, B. J. O’Roak, et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nature Genetics*, 40(5):592–599, 2008.
- [42] C. T. Johansen and R. A. Hegele. Allelic and phenotypic spectrum of plasma triglycerides. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1821(5):833–842, 2012.
- [43] C. T. Johansen, J. Wang, Lanktree, et al. Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genetics*, 42(8):684–687, 2010.
- [44] M. J. Khoury, J. Little, M. Gwinn, and P. A. Ioannidis. On the synthesis and interpretation of consistent but weak gene-disease associations in the era of genome-wide association studies. *International Journal of Epidemiology*, 36(2):439–445, 2007.
- [45] R. Klein, C. Zeiss, E. Y. Chew, et al. Complement factor H polymorphism in age-related macular degeneration. *Science*, 308(5720):385–389, 2005.

- [46] G. V. Kryukov, A. Shpunt, J. A. Stamatoyannopoulos, and S. R. Sunyaev. Power of deep, all-exon resequencing for discovery of human trait genes. *Proceedings of the National Academy of Sciences*, 106(10):3871, 2009.
- [47] C. S. Ku, D. N. Cooper, C. Polychronakos, N. Naidoo, M. Wu, and R. Soong. Exome sequencing: Dual role as a discovery and diagnostic tool. *Annals of Neurology*, 71(1):5–14, 2012.
- [48] C. S. Ku, N. Naidoo, and Y. Pawitan. Revisiting Mendelian disorders through exome sequencing. *Human Genetics*, 129(4):351–370, 2011.
- [49] E. S. Lander. The new genomics: global views of biology. *Science*, 274(5287):536–539, 1996.
- [50] E. S. Lander and D. Botstein. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121(1):185–199, 1989.
- [51] M. B. Lanktree, R. A. Hegele, N. J. Schork, and J. D. Spence. Extremes of unexplained variation as a phenotype. *Circulation: Cardiovascular Genetics*, 3(2):215–221, 2010.
- [52] S. Lee, M. J. Emond, M. Bamshad, K. Barnes, M. Rieder, D. Nickerson, D. Christiani, M. Wurfel, and X. Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, 91:224–237, 2012.
- [53] B. Li and S. M. Leal. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *The American Journal of Human Genetics*, 83:311–321, 2008.
- [54] D. Li, J. P. Lewinger, W. J. Gauderman, C. E. Murcray, and D. Conti. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic Epidemiology*, 35:790–799, 2011.
- [55] D. J. Liu, R. C. Banuelos, and S. M. Leal. A comprehensive evaluation of methods for detecting and interpreting rare variant quantitative trait associations via sequence data. Unpublished paper.
- [56] D. J. Liu and S. M. Leal. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genetics*, 6(10):e1001156, 2010.
- [57] J. R. Lupski, J. W. Belmont, E. Boerwinkle, and R. A. Gibbs. Clan genomics and the complex architecture of human disease. *Cell*, 147(1):32–43, 2011.

- [58] A. J. MacGregor, H. Sneider, N. J. Schork, and T. D. Spector. Twins: novel uses to study complex traits and genetic diseases. *Trends in Genetics*, 16(3):131–134, 2000.
- [59] B. E. Madsen and S. R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2):e1000384, 2009.
- [60] B. Maher. The case of the missing heritability. *Nature*, 456(6):18–21, 2008.
- [61] T. A. Manolio, L. D. Brooks, and F. S. Collins. A HapMap harvest of insights into the genetics of common disease. *The Journal of Clinical Investigation*, 118(5):1590–1605, 2008.
- [62] T. A. Manolio, F. S. Collins, N. J. Cox, et al. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.
- [63] A. Maxmen. Exome sequencing deciphers rare diseases. *Cell*, 144(5):635–637, 2011.
- [64] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.
- [65] McKusick-Nathans Institute of Genetic Medicine. *Online Mendelian Inheritance in Man (OMIM)*. Available at: <http://www.omim.org>, Accessed April 16, 2013.
- [66] S. Morgenthaler and W. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1):28–56, 2007.
- [67] A. Morris and E. Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2):188–193, 2010.
- [68] S. L. Murphy, J. Q. Xu, and K. D. Kochanek. Deaths: Preliminary data for 2010. *National Vital Statistics Report*, 60(4), 2012.
- [69] B. M. Neale, M. A. Rivas, B. F. Voight, et al. Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3):e1001322, 2011.
- [70] S. Nejentsev, N. Walker, D. Riches, M. Egholm, and J. A. Todd. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324(5925):387–389, 2009.

- [71] M. R. Nelson, D. Wegmann, M. G. Ehm, et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 2012.
- [72] P. C. Ng, S. Levy, J. Huang, et al. Genetic variation in an individual human exome. *PLoS Genetics*, 4(8):e1000160, 2008.
- [73] S. B. Ng, A. W. Bigham, K. J. Buckingham, et al. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. *Nature Genetics*, 42(9):790–793, 2010.
- [74] S. B. Ng, K. J. Buckingham, C. Lee, et al. Exome sequencing identifies the cause of a Mendelian disorder. *Nature Genetics*, 42(1):30–35, 2009.
- [75] S. B. Ng, E. H. Turner, P. D. Robertson, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–276, 2009.
- [76] NIH-National Human Genome Research Institute (NHGRI). *Talking Glossary of Genetic Terms*. Available at: <http://www.genome.gov/Glossary/index.cfm?p=viewimage&id=61>, Accessed July 24, 2010.
- [77] NIH Office of Rare Diseases Research. *Undiagnosed Disease Program*. Available at: <http://rarediseases.info.nih.gov/Resources.aspx?PageID=31>, Accessed September 7, 2012.
- [78] B. J. ORoak, L. Vives, S. Girirajan, et al. Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature*, 485.
- [79] J. Ott. *Analysis of human genetic linkage*. Johns Hopkins University Press, Baltimore, 3rd edition, 1999.
- [80] Oxford Nanopore Technologies. *The GridION system*. Available at: <http://www.nanoporetech.com>, Accessed July 23, 2012.
- [81] Pacific Biosciences. *Pacbio RS*. Available at: <http://www.pacificbiosciences.com>, Accessed July 16, 2012.
- [82] L. Peltonen, M. Perola, J. Naukkarinen, and A. Palotie. Lessons from studying monogenic disease for common disease. *Human Molecular Genetics*, 15(1):R67–R74, 2006.
- [83] B. Peng and X. Liu. Simulating sequences of the human genome with rare variants. *Human Heredity*, 70(4):287–291, 2011.
- [84] E. Pennisi. Will computers crash genomics? *Science*, 331(6018):666–668, 2011.



- [85] R. Plomin, C. M. Haworth, and O. S. Davis. Common disorders are quantitative traits. *Nature Review Genetics*, 10:872–878, 2009.
- [86] A. L. Price, G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples, L. J. Wei, and S. R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838, 2010.
- [87] J. K. Pritchard. Are rare variants responsible for susceptibility to complex diseases? *American Journal of Human Genetics*, 69(1):124–137, 2001.
- [88] B. Rabbani, N. Mahdih, K. Hosomichi, H. Nakaoka, and I. Inoue. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *Journal of Human Genetics*, 57:621–632, 2012.
- [89] J. E. Richards and R. S. Hawley. *The Human Genome: A User’s Guide*. Elsevier, San Diego, 3rd edition, 2011.
- [90] C. Rödelberger, P. Krawitz, S. Bauer, J. Hecht, A. W. Bigam, M. Bamshad, B. J. de Condor, M. R. Schweiger, and P. N. Robinson. Identity-by-descent filtering of exome sequence data for disease–gene identification in autosomal recessive disorders. *Bioinformatics*, 27(6):829–836, 2011.
- [91] S. Romeo, L. A. Pennacchio, Y. Fu, et al. Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL. *Nature Genetics*, 39(4):513–516, 2007.
- [92] F. S. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.
- [93] A. M. Saunders, W. J. Strittmatter, D. Schmechel, et al. Association of apolipoprotein E allele  $\epsilon 4$  with late-onset familial and sporadic Alzheimer’s disease. *Neurology*, 43(8):1467–1467, 1993.
- [94] J. Shendure and H. Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008.
- [95] G. Shi and D. C. Rao. Optimum designs for next-generation sequencing to discover rare variants for common complex disease. *Genetic Epidemiology*, 35(6):572–579, 2011.
- [96] P. L. Ståhl and J. Lundeberg. Toward the single-hour high-quality genome. *Annual Review of Biochemistry*, 81(1):359–378, 2012.

- [97] A. Tenesa, S. A. Knott, A. D. Carothers, and P. M. Visscher. Power of linkage disequilibrium mapping to detect a quantitative trait locus (QTL) in selected samples of unrelated individuals. *Annals of Human Genetics*, 67(6):557–566, 2003.
- [98] J. A. Tennessen, T. D. O’Connor, M. J. Bamshad, and J. M. Akey. The promise and limitations of population exomics for human evolution studies. *Genome Biology*, 12(9):127, 2011.
- [99] The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.
- [100] The Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713–720, 2010.
- [101] K. A. Wetterstrand. *DNA Sequencing Cost: Data from the NHGRI Large-Scale Genome Sequencing Program*. Available at: <http://www.genome.gov/sequencingcosts>, Accessed December 14, 2012.
- [102] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89:82–93, 2011.
- [103] B. Xu, J. L. Roos, P. Dexheimer, B. Boone, B. Plummer, S. Levy, J. A. Gogos, and M. Karayiorgou. Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. *Nature Genetics*, 43(9):864–868, 2011.
- [104] J. Yang, B. Benyamin, B. P. McEvoy, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, 2010.
- [105] G. Zhang, D. W. Nebert, R. Chakraborty, and L. Jin. Statistical power of association using the extreme discordant phenotype design. *Pharmacogenetics and Genomics*, 16(6):401–413, 2006.